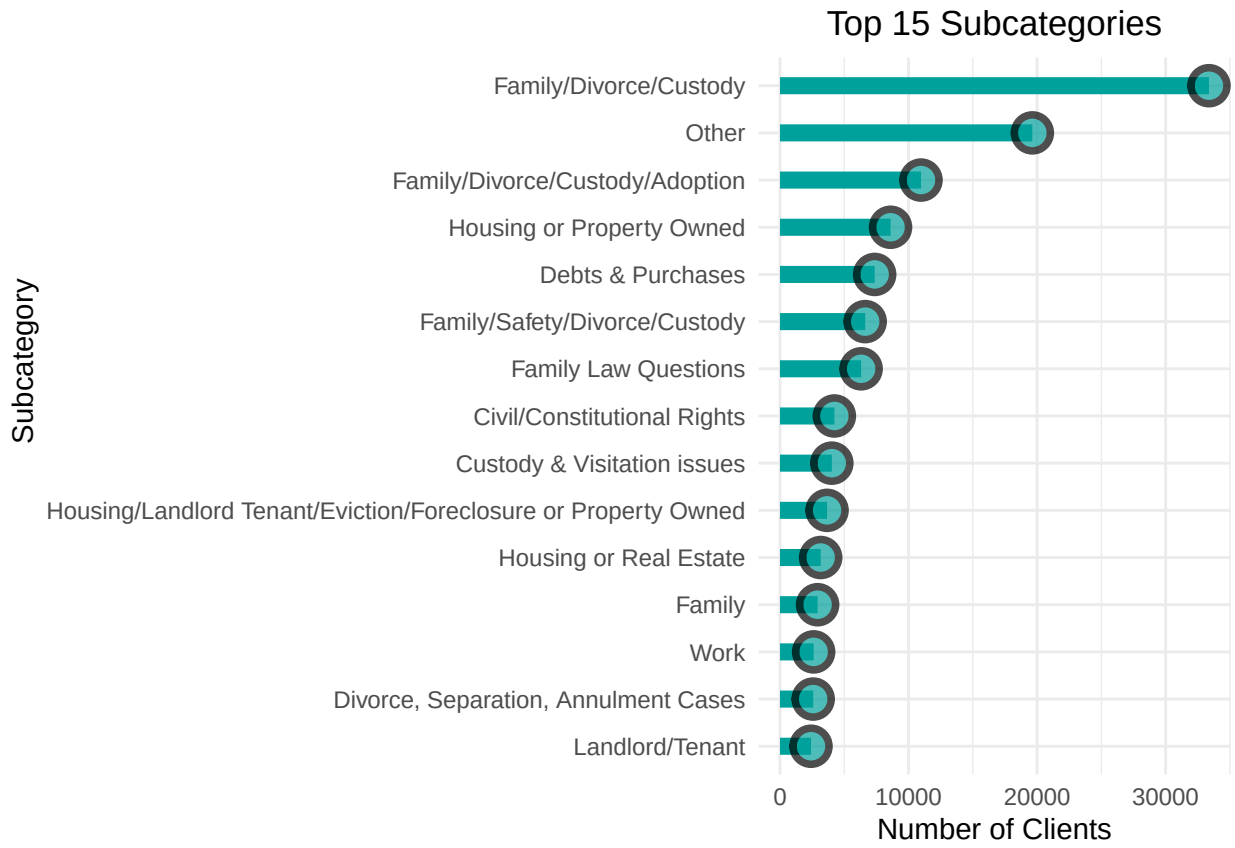
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions



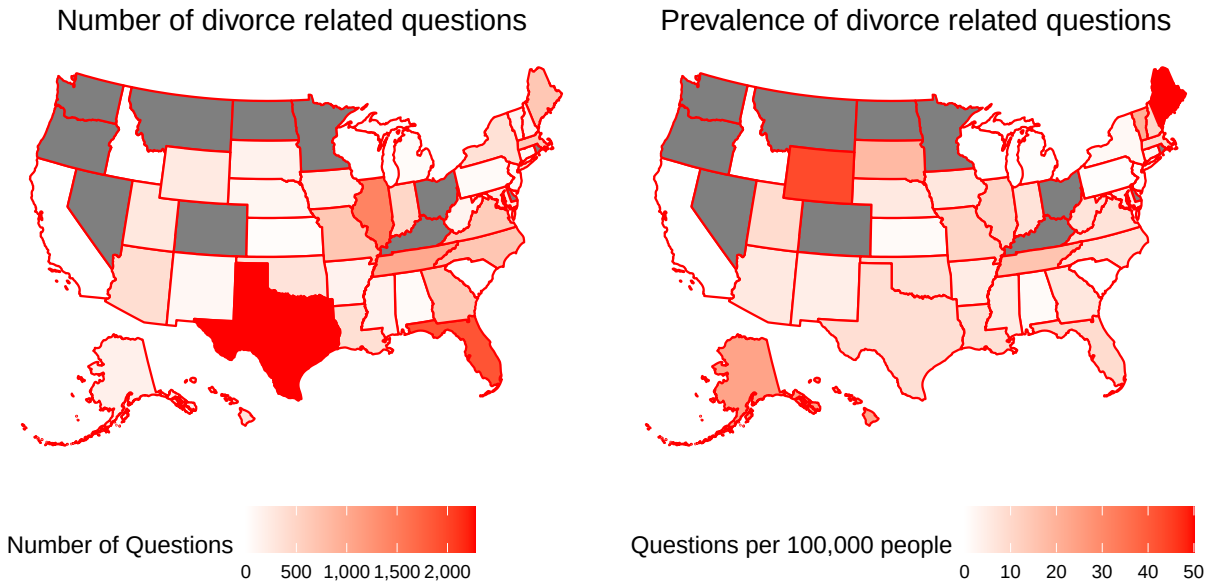
Interpretation of visualization 1:

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



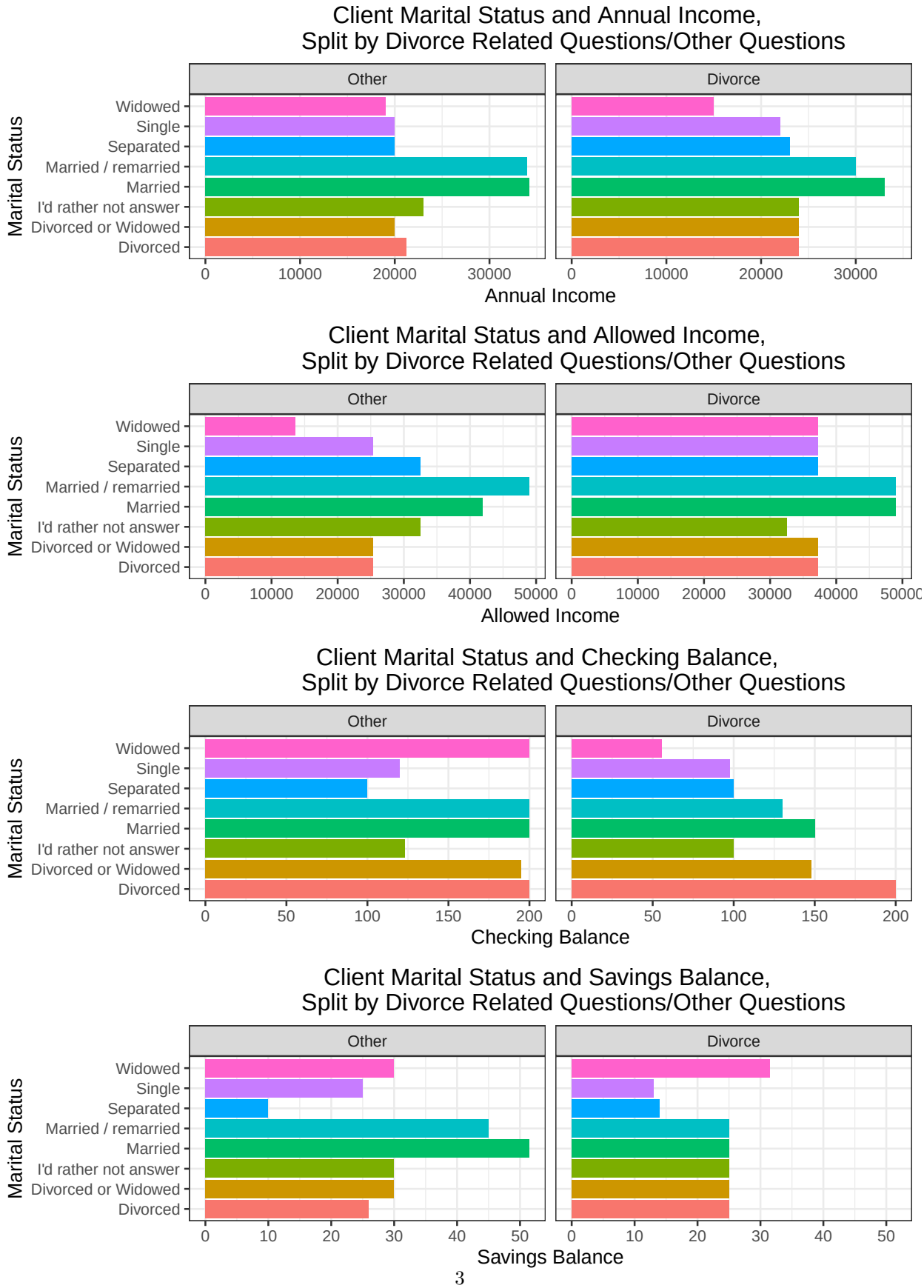
Interpretation of visualization 2:

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 Divorce Rate



Interpretation of visualization 3:

From the above visualization, we can see that the annual income of clients asking divorce-related questions is higher on average compared to other clients. The exception is in the married/remarried category, which itself exhibits higher annual income than other marital statuses across both question categories. The states also recognize this discrepancy and generally allow higher income married individuals to ask pro bono questions. That is, the allowed income is noticeably higher for married or remarried people compared to other marital statuses. If someone wishes to ask a divorce-related question, we can also see that if they do not wish to disclose their marital status, their allowed income to ask a question free of charge drops.

Both the checking and savings balances of clients pursuing divorce-related advice is lower on average than the balances of clients asking other questions. Among all marital statuses asking divorce questions, divorced clients have the highest checking balance and widowed clients have the lowest. This is interesting because both statuses assume a person has no current partner. Clients with similar statuses, like single or separated, have average checking balances in between widowed and divorced clients. However, single and separated clients tend to not have high balances in their savings accounts, no matter what legal advice they are seeking. Unlike the data of annual/allowed income, married clients do not maintain significantly higher checking or savings balances compared to other clients, with the exception of the savings balance of clients asking non divorce-related legal questions.

Research question: What variables are significant predictors of whether a legal question is related to divorce?

Checking for Multicollinearity

When conducting a test for multicollinearity we found that two of our independent variables were highly correlated, with GVIF values much greater than 5, and needed to be removed.

```
##                          GVIF Df GVIF^(1/(2*Df))
## factor(StateAbbr)     1.088084 39        1.001083
## Age                   1.311192  1        1.145073
## NumberInHousehold     1.236508  1        1.111984
## factor(MaritalStatus) 1.530012  7        1.030843
## AnnualIncome          1.166916  1        1.080239
## SavingsBalance        1.092866  1        1.045402
## CheckingBalance       1.094757  1        1.046307
```

Fixing Multicollinearity

After removing the independent variable "AllowedIncome" we find that our multicollinearity is below 5 and therefore reflects the fact that there is an absence of multicollinearity in 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
```

Comparison of Models

```
## $Models
##   Formula
## 1 "factor(Subcategory) ~ (StateAbbr + Age + NumberInHousehold + MaritalStatus + AnnualIncome + Savin
## 2 "Subcategory ~ factor(MaritalStatus) + SavingsBalance"
## 3 "factor(Subcategory) ~ SavingsBalance + CheckingBalance + factor(MaritalStatus) + NumberInHouseho
## 4 "factor(Subcategory) ~ AllowedIncome"
## 5 "factor(Subcategory) ~ factor(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
```

Odds ratio and confidence intervals for the Best Model

```
##                                          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

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

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

COMMENTS

WHAT WE NEED TO FINISH:

- Explain our assumptions (ie. multicollinearity)
- Explain what we intend to achieve with proposed question
- Comment on our findings related to the research question
- clearly state our dependent and independent variables
- Explain why we selected the best model (using lowest AIC, BIC)
- Explain our log odds
- Explain Confidence interval

- Discuss the positive predictor and negative predictor variables (in reference to FL and Single)
- 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)
- SIGN THE FORM SO WE'RE GRADED!!!