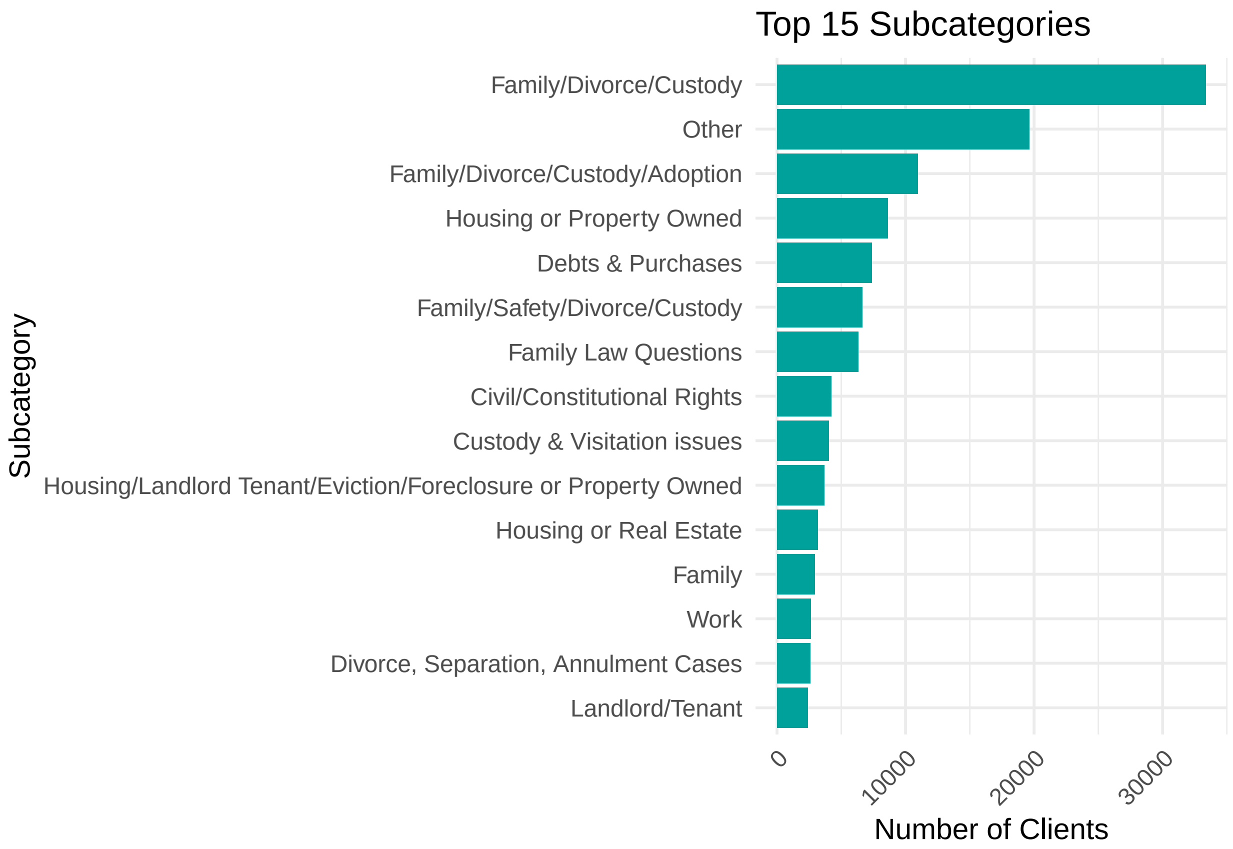
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions

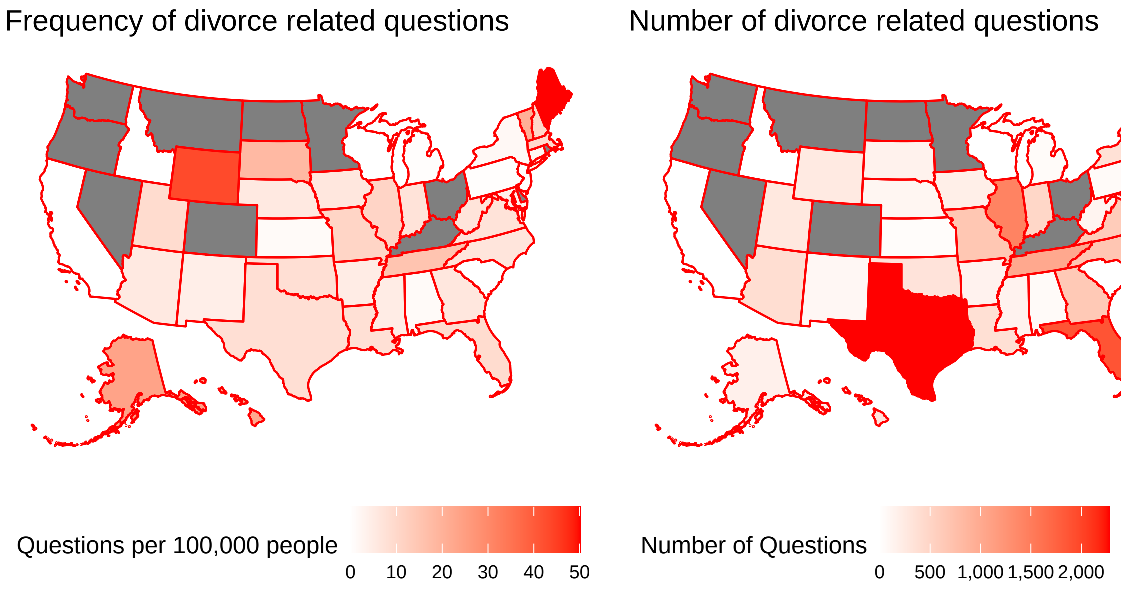


Interpretation:
For our project, we want to started off with investigating the most frequently asked legal questions subcategories so we can define a most common question type based on the data set we have at hand and use it as the theme of our research question.

Based on our side way bar chart, it is clear to see that out of the top 15 subcategories, among the top three categories of questions asked by clients on the online platform, two of them are related to divorce.The question of subcategory Family/Divorce/Custody has the highest frequency among the top 15, and it has been asked by almost twice as many clients than the second subcategory of others, which is showing that the divorce related questions should be treated with more preparation when it comes to training volunteers.

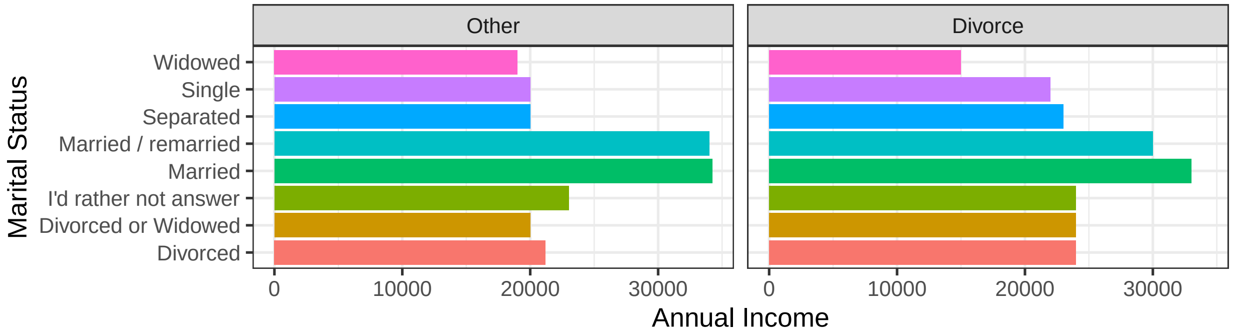
This is intriguing to us as it begs for answers to the questions such as: “What are the major determinants of divorce?”,“Do the clients’ backgrounds affect their tendency to ask divorce related question on the ABA online platform?", and most importantly, "How do we prepare our volunteers to address these divorce related questions?"

Visualization 2: US Map of Divorce Related Questions Distribution

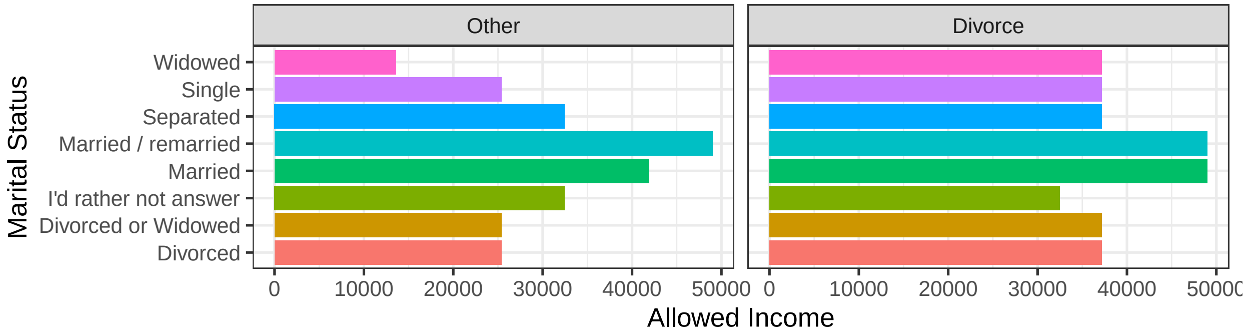

Frequency of divorce related questions
Number of divorce related questions
Questions per 100,000 people
0
10
20
30
40
50
Number of Questions
0
500
1,000
1,500
2,000

Visualization 3: Financial Status Correlation with Divorce Rate

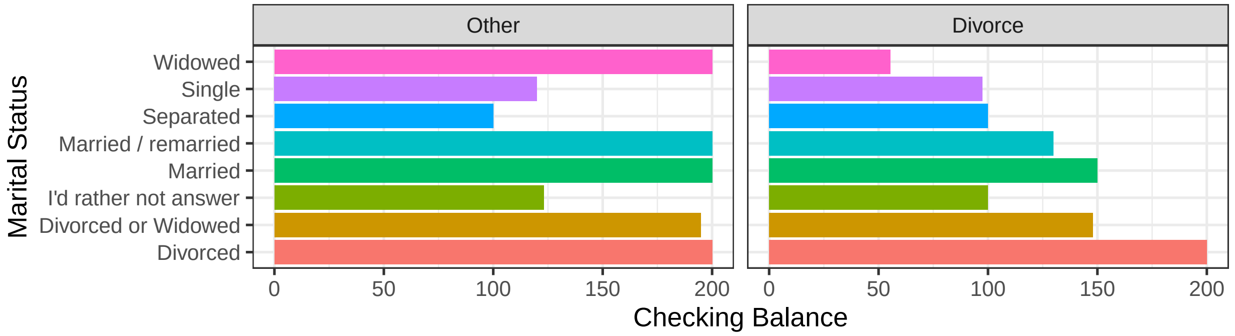
Client Marital Status and Annual Income,
Split by Divorce Related Questions/Other Questions



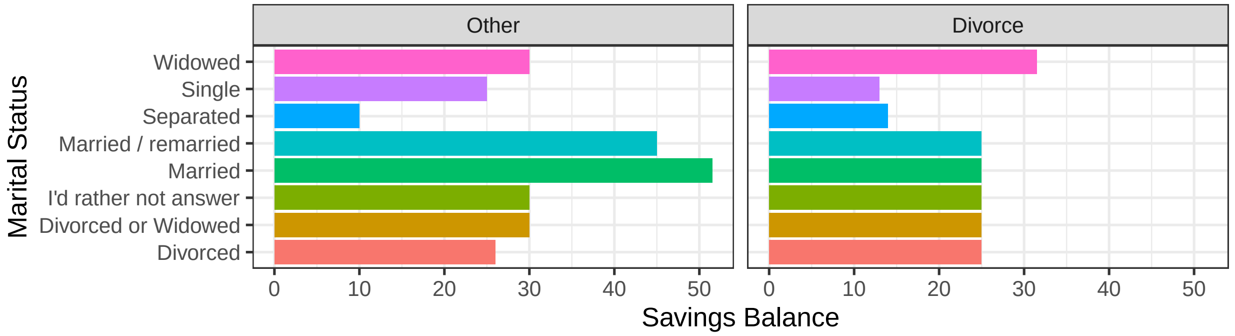
Client Marital Status and Allowed Income,
Split by Divorce Related Questions/Other Questions



Client Marital Status and Checking Balance,
Split by Divorce Related Questions/Other Questions



Client Marital Status and Savings Balance,
Split by Divorce Related Questions/Other Questions



Should we factor the categorical variables like StateAbbr and MaritialStatus for model1??

```
## $Models
##   Formula
## 1 "factor(Subcategory) ~ factor(StateAbbr) + Age + NumberInHousehold + factor(MaritalStatus) + Annual
## 2 "Subcategory ~ factor(MaritalStatus) + SavingsBalance"
## 3 "factor(Subcategory) ~ SavingsBalance + CheckingBalance + factor(MaritalStatus) + NumberInHousehold
## 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   53  50570 49540 49540 50020 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
```

Model Accuracy

```
##
##     FALSE  TRUE
##   0 32077  3186
##   1  9593  5764
```

```
## [1] "The accuracy of our optimum model was: 74.76%."
```

COMMENTS

Daniel: - It appears that our model 1 is the most accurate since it has the lowest AIC. - I think we can only use 1 model for viz 2 and viz 3, otherwise our pdf page count is more than 8 pages. - For viz 2, it might be better to look at the number of divorces per the population because as we'd expect, states with a large population would have more divorce related questions.

- Also a bit weird that state like California and Idaho don't have any divorce related questions. I think thats because when we filtered the subcategories, those states don't have any subcategories with "Divorce". I don't know how we'd account for that… but we might need to so Prince doesn't mark us off.

Daniel: - We might need to remove the StateAbbr "US"??? Billy: Prolly - Actually we might not need to remove StateAbbr "US". Looking at the instructions PDF, is says that the "US" is a category for immigrants and veterans. So while its not a state in itself, still might be important to keep??

Comment by Billy: - The AIC is alarmingly high - Some assumptions to consider for our model - observations are independent - no severe multicollinearity among explanatory variables, no highly correlated variables - no extreme outliers - exist linear relationship between explanatory variable and logit of response linearly related to log odds - Large sample set - Remove outliers - Remove independent variables with collinealarity

Datasets: - Statesites - StateAbbr - AllowedAssets - BaseIncomeLimit - PerHouseholdMemberIncomeLimit - IncomeMultiplier - Client - StateAbbr - Prettymuch the entire dataset - Categories - Category: Family and Children -nSubcategories: everything related to divorces

Comment by Billy: - The AIC is alarmingly high - Some assumptions to consider for our model - observations are independent - no sever multicollinearity among explanatory variables, no highly correlated variables - no extreme outliers - exist linear relationship between explanatory variable and logit of response linearly related to log odds - Large sample set - Remove outliers - Remove independent variables with collinealarity

Questions: You were saying that we only need one of the residual analysis - logit model, the default should be binomial - interpret log log odds, don't have to interpret odds ratio - Confidence interval - Explaining why we use AIC VIC - Check accuracy of model - Do the residual analysis to check for accuracy only for our best model. - Suggestion based on limitations , missing data = limitation - 1 summary for 1 model we submitted at the end

Summary: come up with a story for our regression