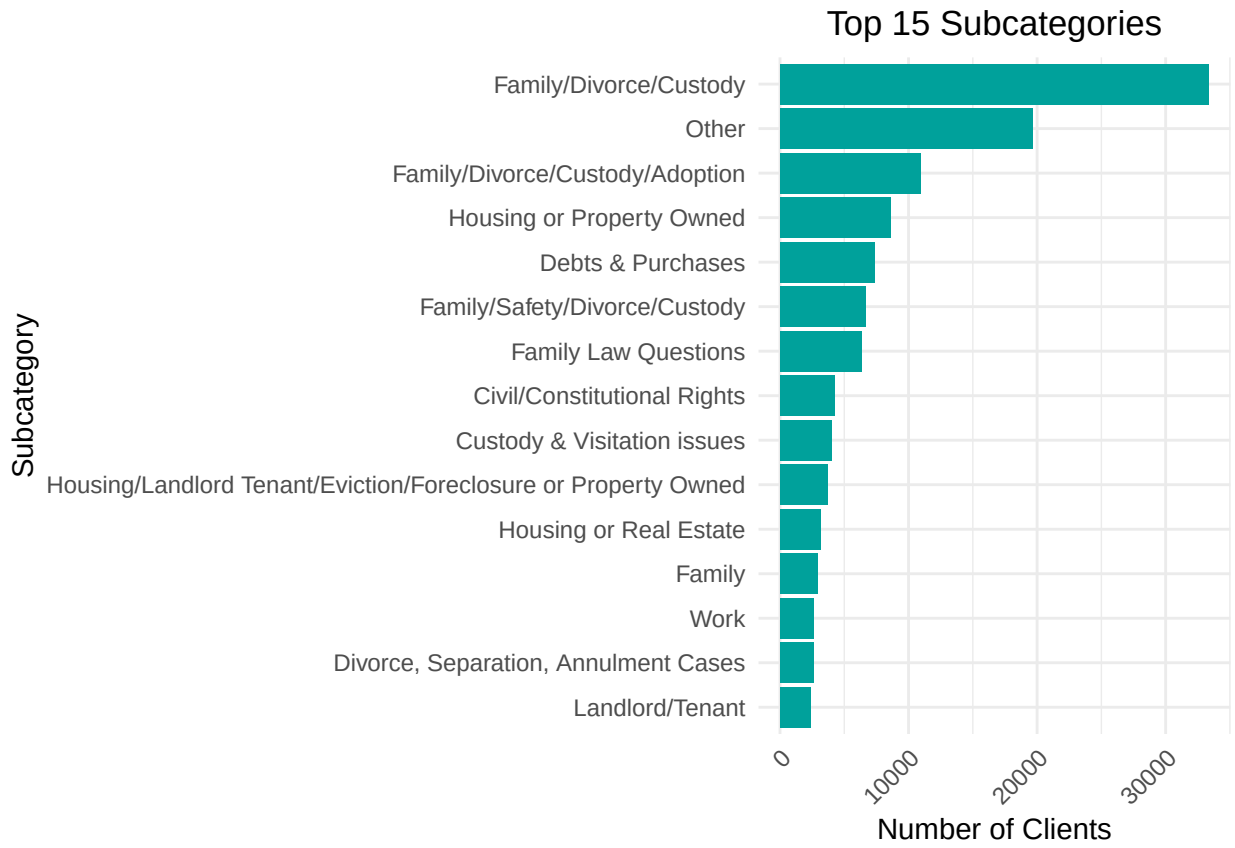
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions



Interpretation:

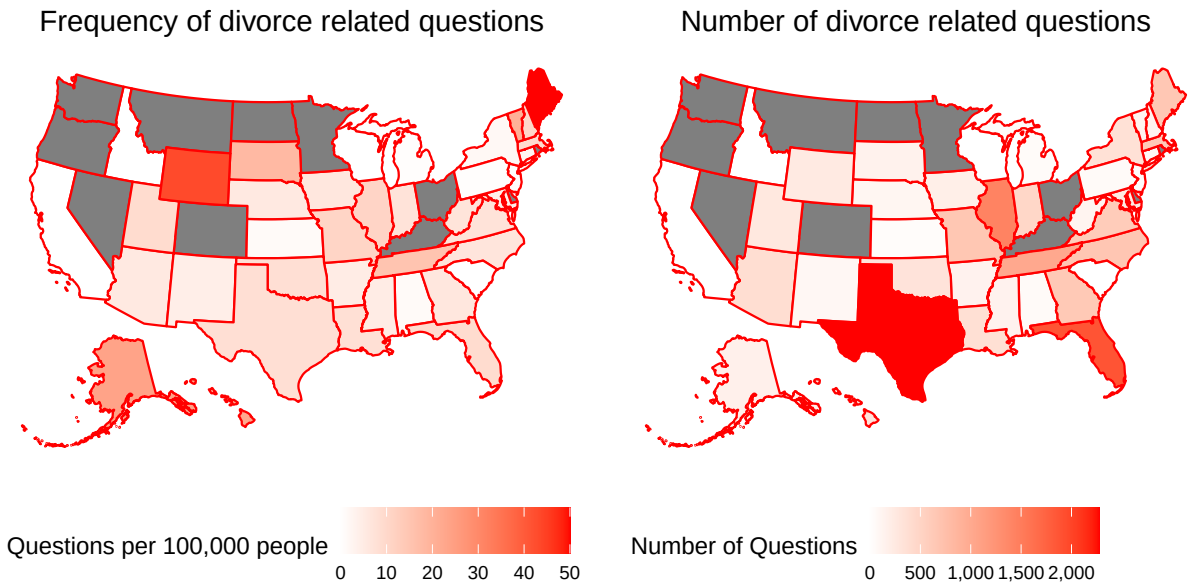
For our project, we want to started off with investigating the most frequently asked legal questions subcategories so we can define a most common question type based on the data set we have at hand and use it as the theme of our research question.

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

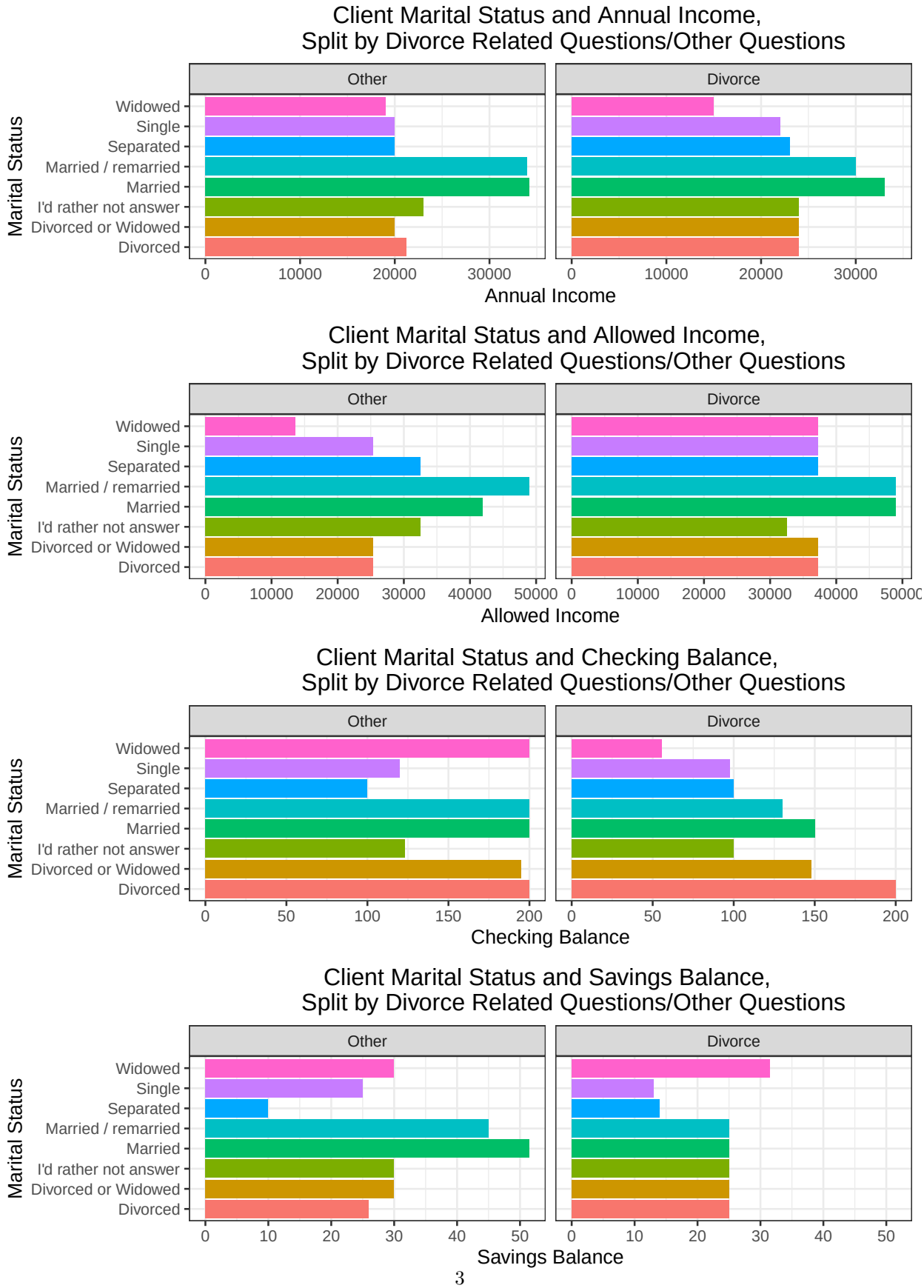


Interpretation:

Certain States such as: Washington, Oregon, Nevada, Montana, Colorado, Minnesota, Ohio, and Kentucky are shown in gray because out of confidentiality and per state law, they are not allowed to provided clients' information, including the category of legal question they ask.

For the 42 states that allow the disclosure of client information, Texas and Florida clearly have the highest number of divorce-related legal questions being asked. However, the distribution of the divorce question frequency does paints a different picture. Texas and Florida no longer stand out when we measure the number of questions asked per 100,000 residents. Here, the states of Wyoming and Maine stand out as places where divorce-related legal advice is most frequently sought on the online platform.

Visualization 3: Financial Status Correlation with Divorce Rate



Interpretation:

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.

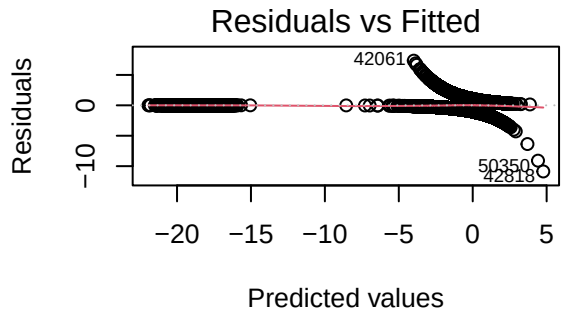
Both the checking and savings balances of clients pursuing divorce-related advice is lower on average than the balances of clients asking other questions. Among all marital statuses asking divorce questions, divorced clients have the highest checking balance and widowed clients have the lowest. This 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.

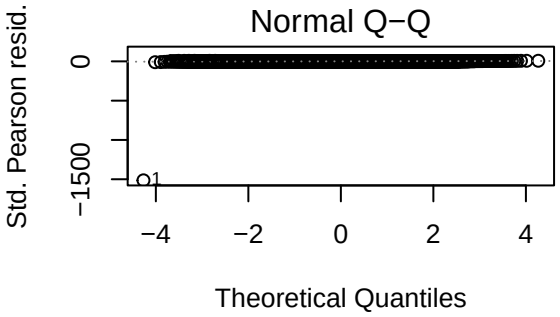
Research question: What variables are significant predictors of whether a legal question is related to divorce?

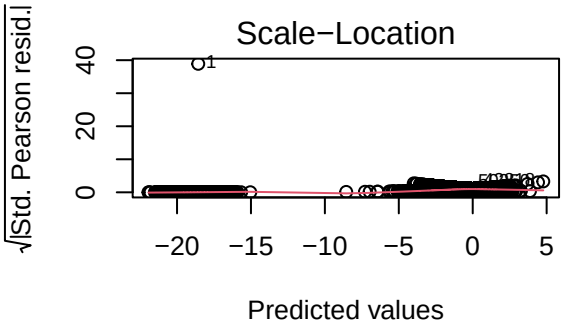
```
# Model 2 Residuals
par(mfrow = c(2, 2))
plot(model1)
```

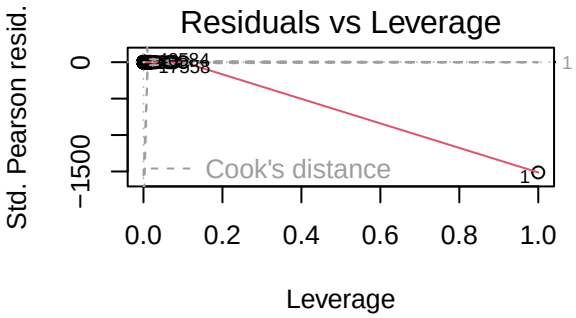
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```










```
vif(model1)
```

```
##                          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
```

```
vif(model1)
```

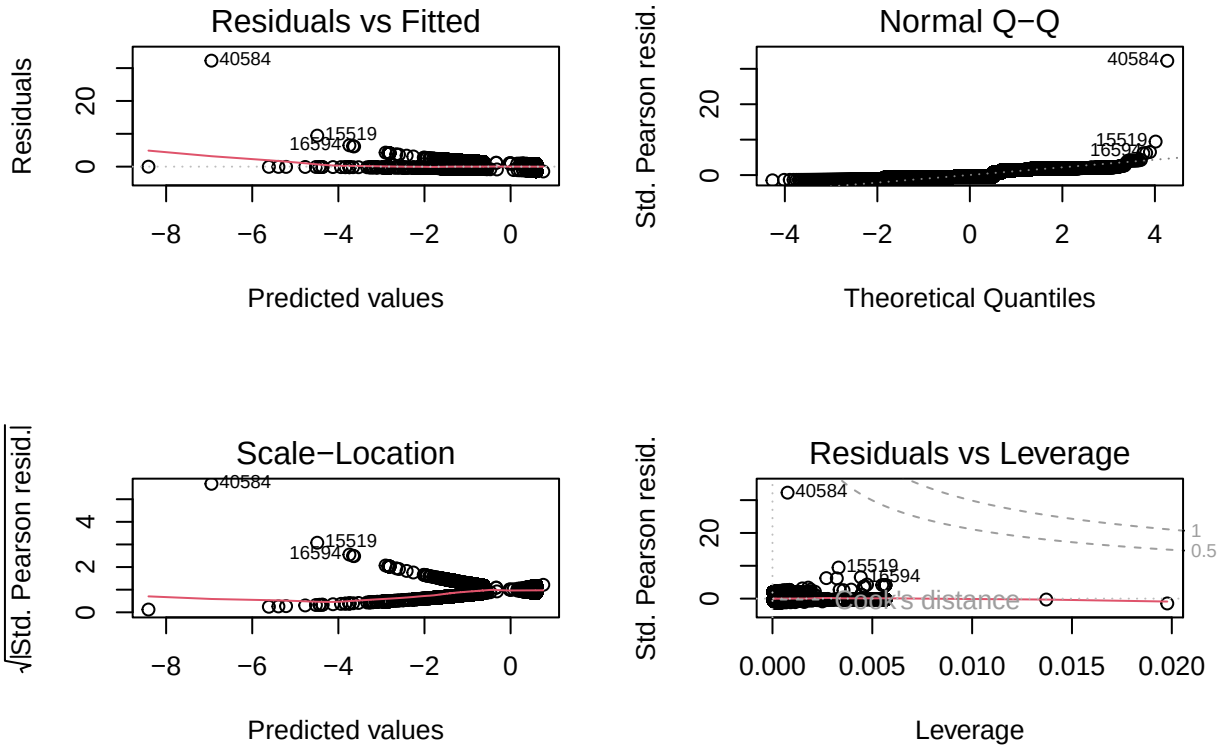
```
##                          GVIF Df GVIF^(1/(2*Df))
## factor(StateAbbr)     1.088084 39        1.001083
## Age                   1.311192  1        1.145073
## NumberInHousehold     1.236508  1        1.111984
## factor(MaritalStatus) 1.530012  7        1.030843
## AnnualIncome          1.166916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

```
# Model 2: Marital Status and Savings Balance influence on divorce related questions
model2 <- glm(Subcategory ~ factor(MaritalStatus) + SavingsBalance ,
```

```
              data = data_overall, family = 'binomial')
summary(model2)
```

```
##
## Call:
## glm(formula = Subcategory ~ factor(MaritalStatus) + SavingsBalance,
##     family = "binomial", data = data_overall)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5118  -0.8878  -0.6863   0.9525   3.7279
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                -8.992e-01  7.051e-02 -12.753
## factor(MaritalStatus)Divorced or Widowed    1.735e-01  7.417e-02   2.339
## factor(MaritalStatus)I'd rather not answer -2.463e-01  9.252e-02  -2.662
## factor(MaritalStatus)Married                2.163e-02  9.071e-02   0.238
## factor(MaritalStatus)Married / remarried    2.664e-01  7.292e-02   3.654
## factor(MaritalStatus)Separated              1.498e+00  7.689e-02  19.479
## factor(MaritalStatus)Single                -4.265e-01  7.253e-02  -5.880
## factor(MaritalStatus)Widowed               -1.914e+00  3.329e-01  -5.751
## SavingsBalance                             -3.111e-05  4.456e-06  -6.982
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## factor(MaritalStatus)Divorced or Widowed   0.019329 *
## factor(MaritalStatus)I'd rather not answer 0.007768 **
## factor(MaritalStatus)Married               0.811500
## factor(MaritalStatus)Married / remarried   0.000258 ***
## factor(MaritalStatus)Separated              < 2e-16 ***
## factor(MaritalStatus)Single                4.09e-09 ***
## factor(MaritalStatus)Widowed               8.88e-09 ***
## SavingsBalance                             2.91e-12 ***
## ---
## 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: 58642  on 50611  degrees of freedom
## AIC: 58660
##
## Number of Fisher Scoring iterations: 5
```

```
# Model 2 Residuals
par(mfrow = c(2, 2))
plot(model2)
```




```
vif(model2)
```

```
##                           GVIF Df GVIF^(1/(2*Df))
## factor(MaritalStatus) 1.001845  7        1.000132
## SavingsBalance        1.001845  1        1.000922
```

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

Model Accuracy

```
##
```

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

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

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: WHAT we need to finish today - The AIC is alarmingly high - Some assumptions to consider for our model - observations are independent Things we have to do: - no sever multicollinearity among explanatory variables, no highly correlated variables - no extreme outliers - Remove independent variables with collinealarity - logit model, the default should be binomial - Explaining why we use AIC BIC - Check accuracy of model - Suggestion based on limitations , missing data = limitation - Confidence interval - Do the residual analysis to check for accuracy only for our best model.