Code

Eric Chen, Junhan Li, & Daniel Fredin

```
# Remove unwanted columns for all datasets
attorneys <- attorneys %>%
  select(-Id, -CreatedUtc, -City, -PostalCode)
attorneys_time <- attorneys_time %>%
  select(-Id, -EnteredOnUtc, -TimeEntryUno)
categories <- categories %>%
  select(-Id)
clients <- clients %>%
  select(-Id, -PostalCode, -CreatedUtc, -InvestmentsBalance)
questionposts <- questionposts %>%
  select(-Id, -CreatedUtc)
questions <- questions %>%
  select(-Id, -AskedOnUtc, -TakenOnUtc, -ClosedOnUtc, -LegalDeadline, -ClosedByAttorneyUno)
statesites <- statesites %>%
  select(-Id)
subcategories <- subcategories %>%
  select(-Id)
```

```
# Convert characters to numeric in client dataset
clients$Age <- as.integer(clients$Age)
```

Data Cleaning

```
## Warning: NAs introduced by coercion
```

```
clients$NumberInHousehold <- as.integer(clients$NumberInHousehold)
```

```
## Warning: NAs introduced by coercion
```

```
clients$AnnualIncome <- as.numeric(clients$AnnualIncome)
```

```
## Warning: NAs introduced by coercion
```

```
clients$AllowedIncome <- as.numeric(clients$AllowedIncome)
```

```
## Warning: NAs introduced by coercion
```

```
clients$CheckingBalance <- as.numeric(clients$CheckingBalance)
```

```
## Warning: NAs introduced by coercion
```

```
clients$SavingsBalance <- as.numeric(clients$SavingsBalance)
```

```
## Warning: NAs introduced by coercion
```

```
# rename columns to match other datasets
questions <- questions %>%
  rename("ClientUno" = "AskedByClientUno", "AttorneyUno" = "TakenByAttorneyUno")
```

```
# Merge into two dataframes
# df1 is all datasets except attorney and attorney_time
df1 <- left_join(clients, questions, join_by("StateAbbr", "ClientUno")) %>%
  left_join(., categories, join_by("StateAbbr", "CategoryUno", "Category")) %>%
  left_join(.,subcategories, join_by("StateAbbr", "CategoryUno", "Subcategory", "SubcategoryUno")) %>%
  left_join(., questionposts, join_by("StateAbbr", "QuestionUno")) %>%
  left_join(., statesites, join_by("StateAbbr", "StateName"))
# df2 is the two attorney datasets
df2 <- left_join(attorneys, attorneys_time, join_by("StateAbbr", "AttorneyUno"))
```

```
# Select the independent variables we want to observe and Remove all NA
data_overall <- df1 %>%
  select(StateAbbr, Age, NumberInHousehold, MaritalStatus, AnnualIncome, SavingsBalance, CheckingBalanc
  distinct(ClientUno, .keep_all = TRUE) %>%
  select(-ClientUno) %>%
  na.omit(df1)
# Remove all NUll from the MaritalStatus variable
data_overall <- data_overall %>%  filter(data_overall$MaritalStatus != "NULL")
```

Change the subcategory to binary categorical data

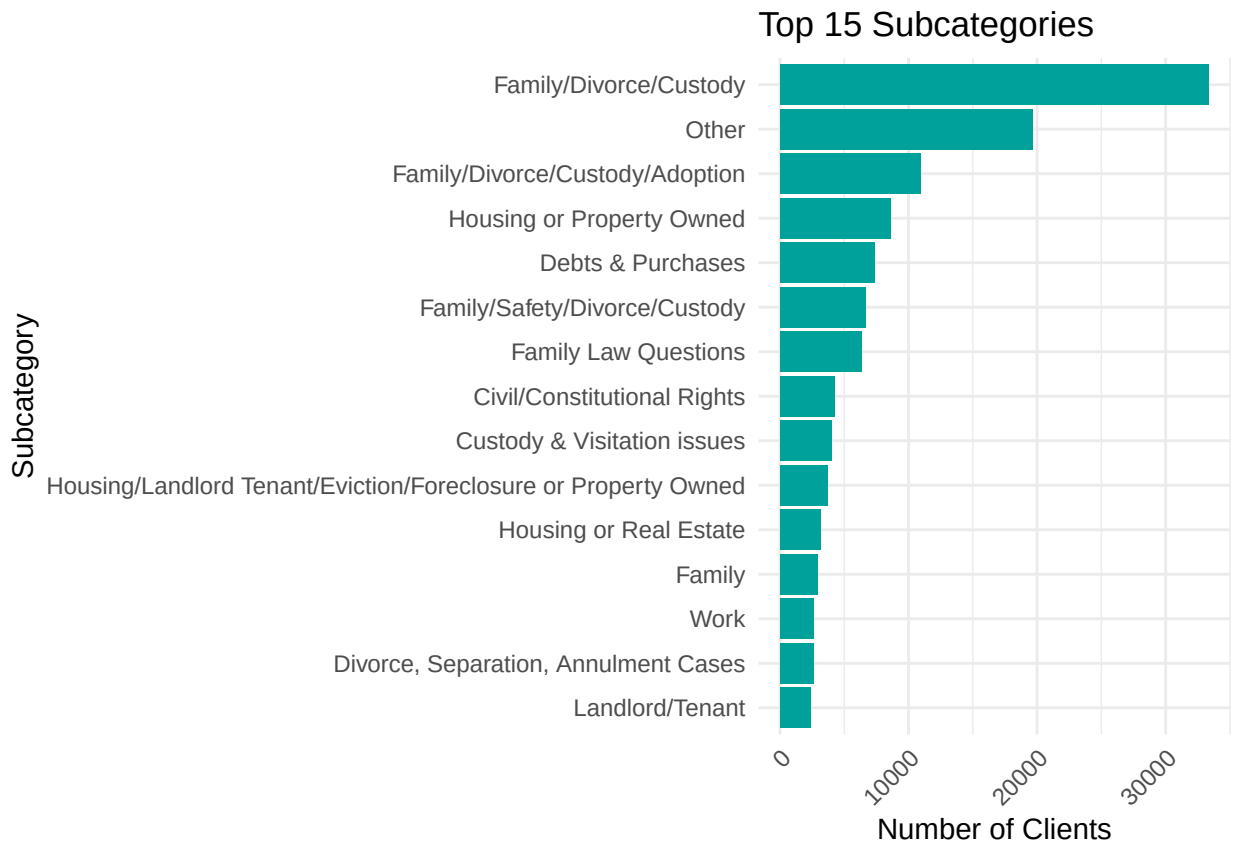
Divorce mentioned entry is represented by 1, otherwise 0.

```
data_overall$Subcategory <- ifelse(str_detect(data_overall$Subcategory, "Divorce"), 1, 0)
```

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

Visualization 1: Investigating the Top 15 Subcategories of Asked Questions

```
top_subcats <- questions %>%
  group_by(Subcategory) %>%
  summarise(num_subcats = n()) %>%
  ungroup()
top_subcats <- top_subcats %>%
  arrange(desc(num_subcats)) %>%
  head(15)
ggplot(top_subcats, aes(x = num_subcats, y = reorder(Subcategory, num_subcats))) +
  geom_bar(stat = "Identity", fill = "#00A19B") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Top 15 Subcategories",
       x = "Number of Clients",
       y = "Subcategory")
```



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.

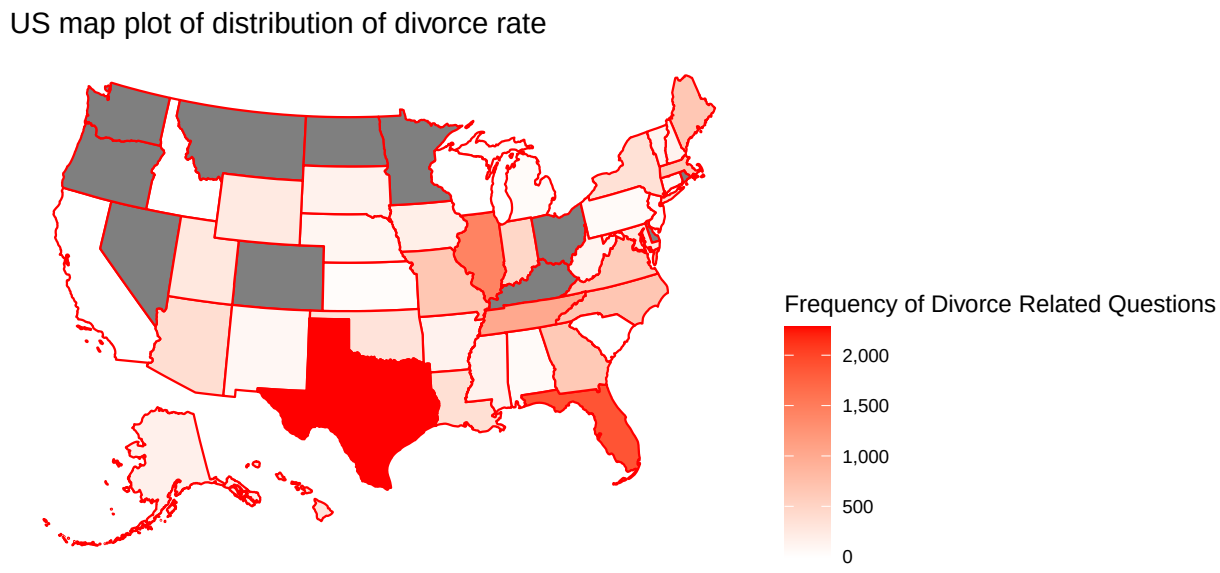
Based on our side way bar chart, it is clear to see that out of the top 15 subcategories, among the top three categories of questions asked by clients on the online platform, two of them are related to divorce.The question of subcategory Family/Divorce/Custody has the highest frequency among the top 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?”,“Does 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

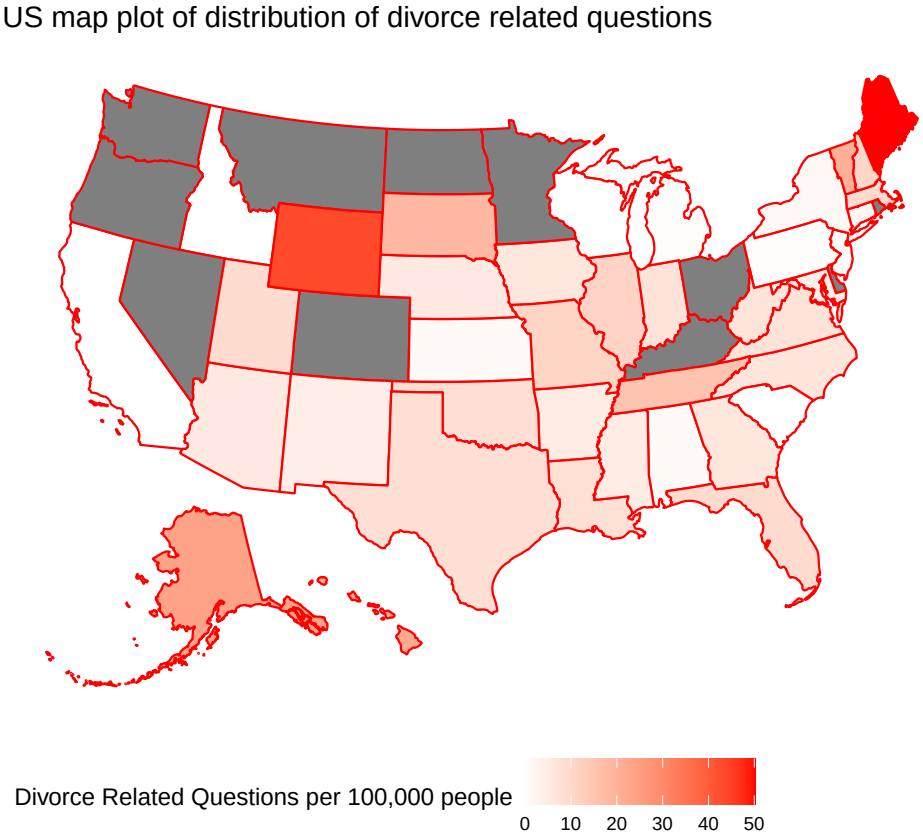
```
# Create a data set with divorce related questions frequency per state
contingency <- table(data_overall$StateAbbr,data_overall$Subcategory)
divorce_distrib <- as.data.frame(contingency)
# Change the column name var1 to state, so we can use plot_usmap
colnames(divorce_distrib)[which(names(divorce_distrib) == "Var1")] <- "state"
# Selects the number of divorces per state
divorce_plot <- divorce_distrib[divorce_distrib$Var2 != 0, ]
# Plots the distribution of divorces in each state
plot_usmap(data = divorce_plot, values = "Freq", color = "red") +
  scale_fill_continuous(
    low = "white", high = "red", name = "Frequency of Divorce Related Questions"
    , label = scales::comma) +
  theme(legend.position = "right") +
  labs(title = "US map plot of distribution of divorce rate")
```



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

However, the US map plot of distribution of divorce questions mentioned in conversation does varied by states. Foe example

```
# Create a data set with divorce related questions frequency per state
contingency <- table(data_overall$StateAbbr,data_overall$Subcategory)
divorce_distrib <- as.data.frame(contingency)
# Change the column name var1 to state, so we can use plot_usmap
colnames(divorce_distrib)[which(names(divorce_distrib) == "Var1")] <- "state"
# Selects the number of divorces per state
divorce_plot <- divorce_distrib[divorce_distrib$Var2 != 0, ]
# Changes state variable back to charcter
divorce_plot$state <- as.character(divorce_plot$state)
# Retrieve state populations and rename abbr column to state
states_pop <- statepop
states_pop <- states_pop %>%
  rename("state" = "abbr")
# Merge the population data with divorce data
dfDivorce <- left_join(states_pop, divorce_plot, join_by("state")) %>%
  select(-fips, -full) %>%
  mutate(distrib = ((Freq/pop_2015)*100000)) %>%
  select(-Freq, -pop_2015)
# Plots the distribution of divorces in each state per population
plot_usmap(data = dfDivorce, values = "distrib", color = "red") +
  scale_fill_continuous(
    low = "white", high = "red", name =
      "Divorce Related Questions per 100,000 people"
    , label = scales::comma) +
  theme(legend.position = "bottom") +
  labs(title = "US map plot of distribution of divorce related questions")
```



This is 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...

Also the project instructions say to use ggplot2 for our visualization. I find that plot_usmap is its own package so should we try to create a us map plot using ggplot2?

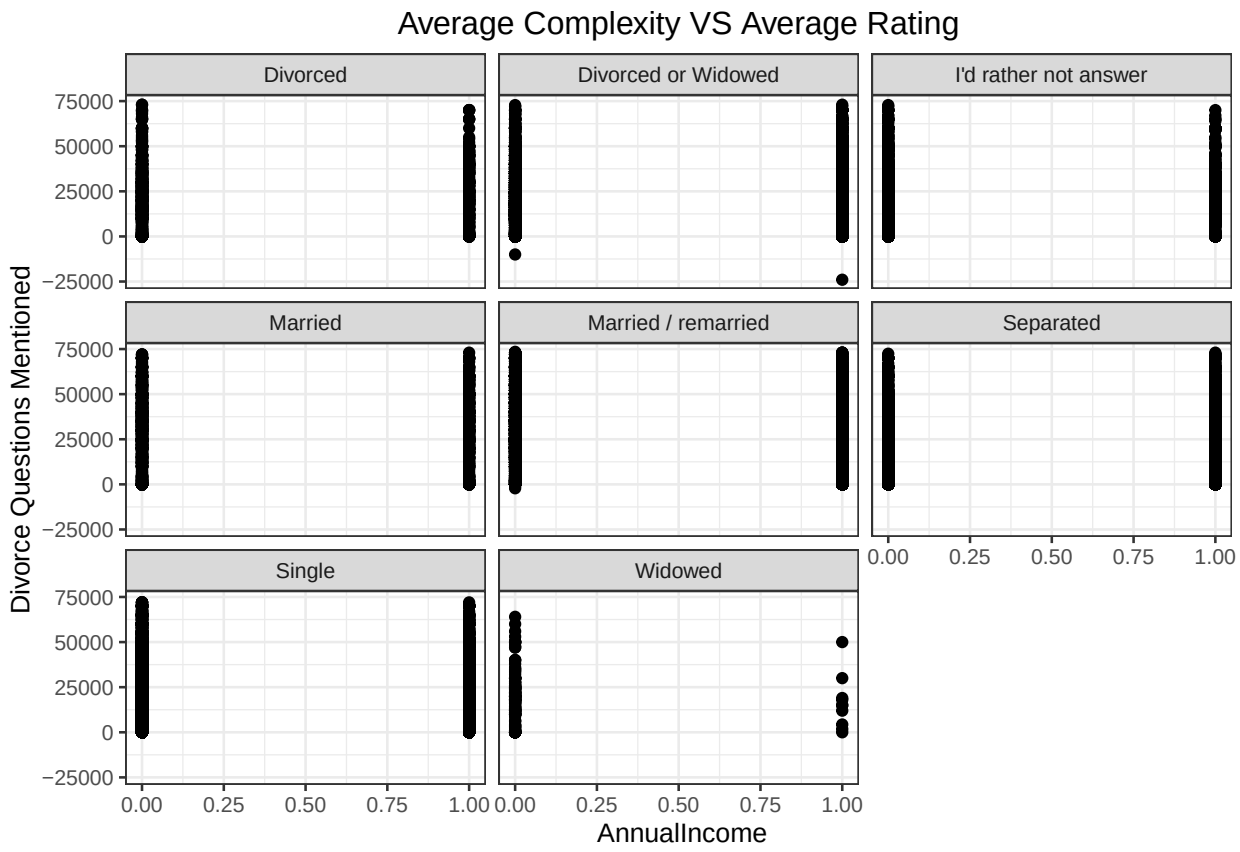
Visualization 3: Financial Status Correlation with Divorce Rate

We need to use that facet, wrap, colored shit here.

```
# Get rid of outliers
quartiles <- quantile(data_overall$AnnualIncome, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(data_overall$AnnualIncome)
Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR
data_noOutlier <- subset(data_overall, AnnualIncome > Lower & AnnualIncome < Upper)

data_noOutlier %>%
ggplot(aes(x = Subcategory , y = AnnualIncome)) +
  geom_point() +
  labs(x = " AnnualIncome",
       y = "Divorce Questions Mentioned",
```

```
      title = "Average Complexity VS Average Rating") +
  facet_wrap(~MaritalStatus) +
  theme_bw(base_size = 10) +
  theme(plot.title =
          element_text(hjust = 0.5))
```




```
data_overall$Subcategory <- ifelse(str_detect(data_overall$Subcategory, "Divorce"), 1, 0)
```

Testing models

Model 1

```
model <- glm(factor(Subcategory) ~ . -Category -StateAbbr, data = data_overall, family = 'binomial')
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(model)
```

```
##
## Call:
## glm(formula = factor(Subcategory) ~ . - Category - StateAbbr,
##     family = "binomial", data = data_overall)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -2.657e+01  1.333e+04  -0.002    0.998
## Age                                5.740e-14  1.285e+02   0.000    1.000
## NumberInHousehold                 -3.847e-14  3.097e+03   0.000    1.000
## MaritalStatusDivorced or Widowed  -9.914e-14  1.196e+04   0.000    1.000
## MaritalStatusI'd rather not answer 4.015e-13  1.454e+04   0.000    1.000
## MaritalStatusMarried               3.593e-13  1.465e+04   0.000    1.000
## MaritalStatusMarried / remarried   4.241e-13  1.184e+04   0.000    1.000
## MaritalStatusSeparated             3.969e-13  1.252e+04   0.000    1.000
## MaritalStatusSingle                1.460e-12  1.171e+04   0.000    1.000
## MaritalStatusWidowed              -7.222e-13  2.877e+04   0.000    1.000
## AnnualIncome                      -1.874e-18  7.879e-02   0.000    1.000
## SavingsBalance                    -2.624e-17  4.239e-01   0.000    1.000
## CheckingBalance                   -5.569e-18  7.609e-01   0.000    1.000
## AllowedIncome                     -1.015e-18  2.199e-01   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.0000e+00  on 50619  degrees of freedom
## Residual deviance: 2.9368e-07  on 50606  degrees of freedom
## AIC: 28
##
## Number of Fisher Scoring iterations: 25
```

Assumes that there is no severe multicollinearity Assume