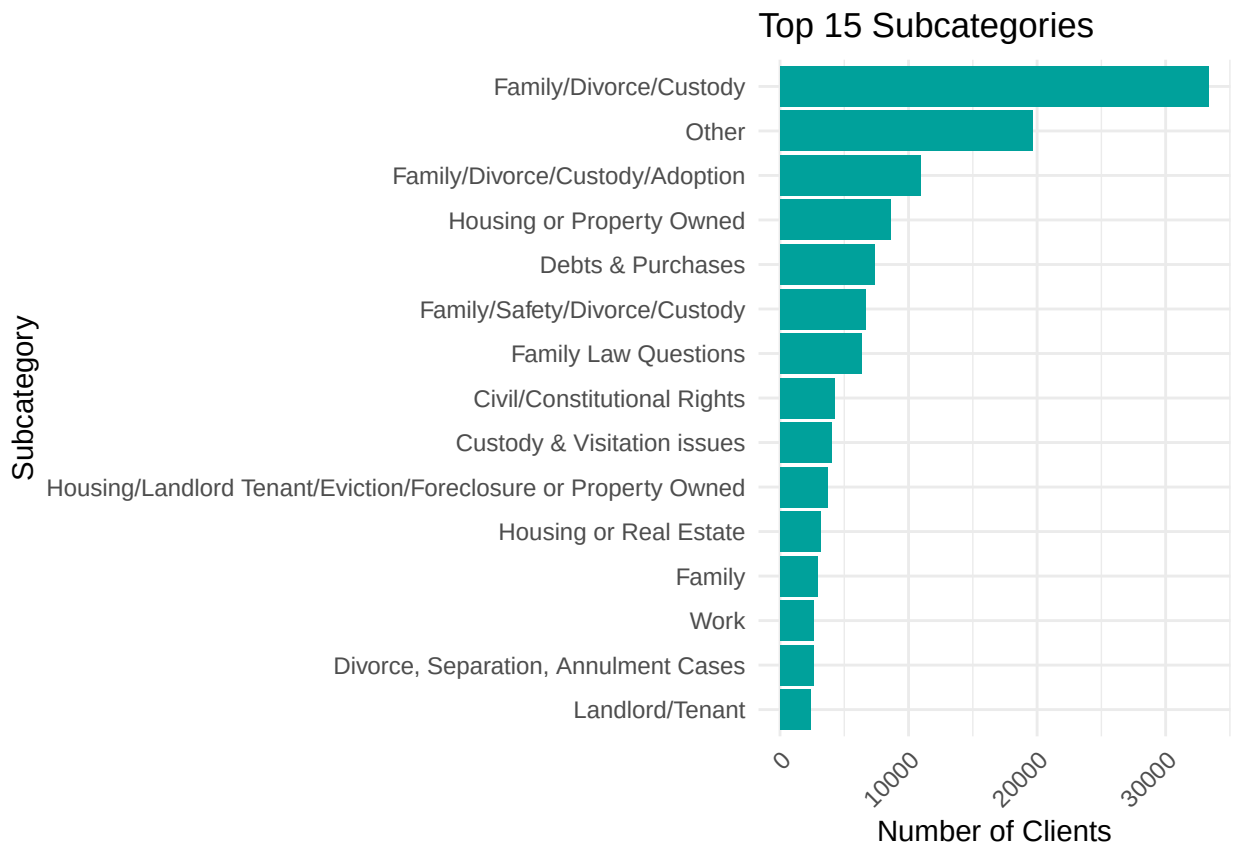
Project 1 - Analysis of American Bar Association data

Eric Chen, Junhan Li, & Daniel Fredin

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

```
# Visualization 1
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.

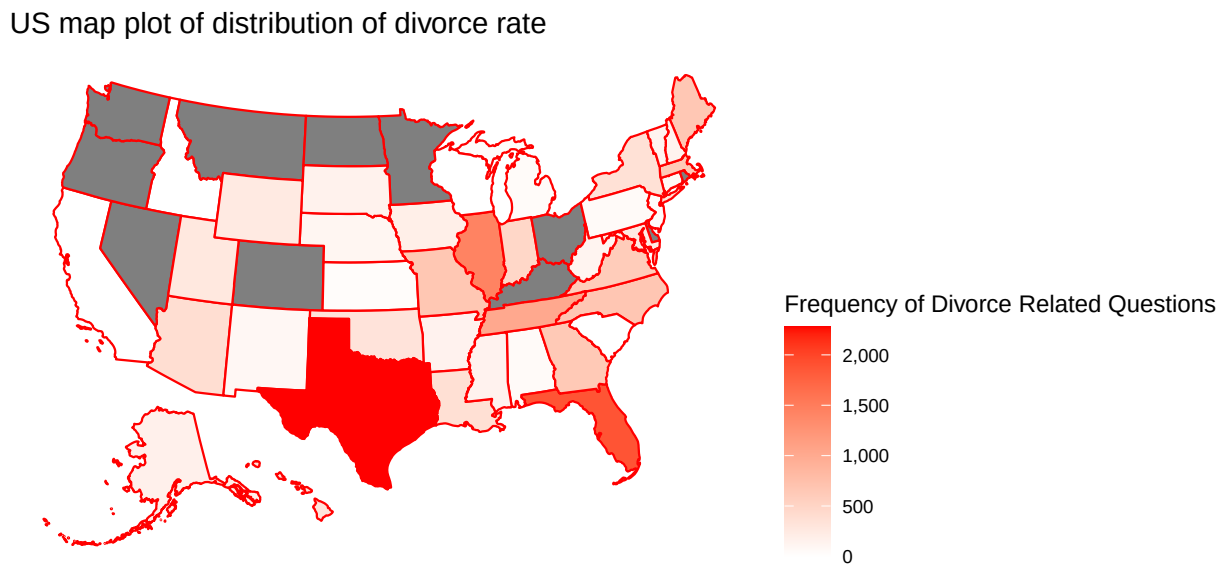
Based on our side way bar chart, it is clear to see that out of the top 15 subcategories, among the top three categories of questions asked by clients on the online platform, two of them are related to divorce.The question of subcategory Family/Divorce/Custody has the highest frequency among the top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

```
# Visualization 2
# 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")
```

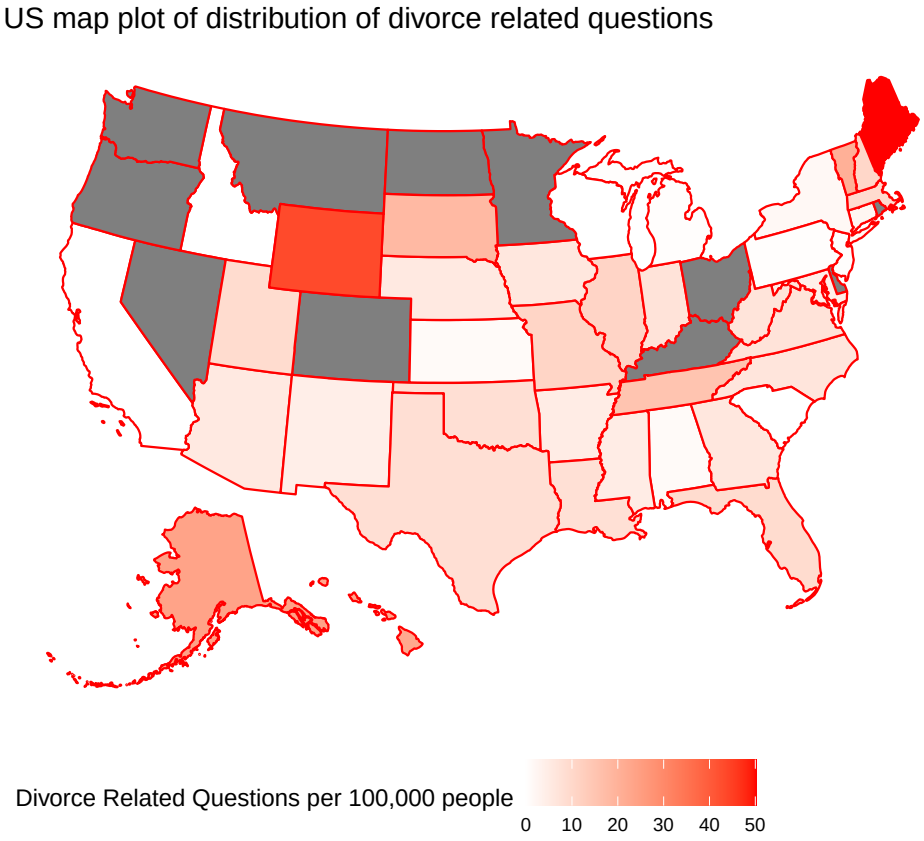


Certain States such as: Washington, Oregon, Nevada, Montana, Colorado, Minnesota, Ohio, and Kentucky are shown in gray because out of confidentiality and per state law, they are not allowed to provided clients' information.

However, the distribution of the divorce questions frequency does varied by states. For example, states such as Texas and Florida has the highest frequency of divorce related questions discuss with the attorney on the online platform.

```
# Visualization 2
# 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)
```

```r
# 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...

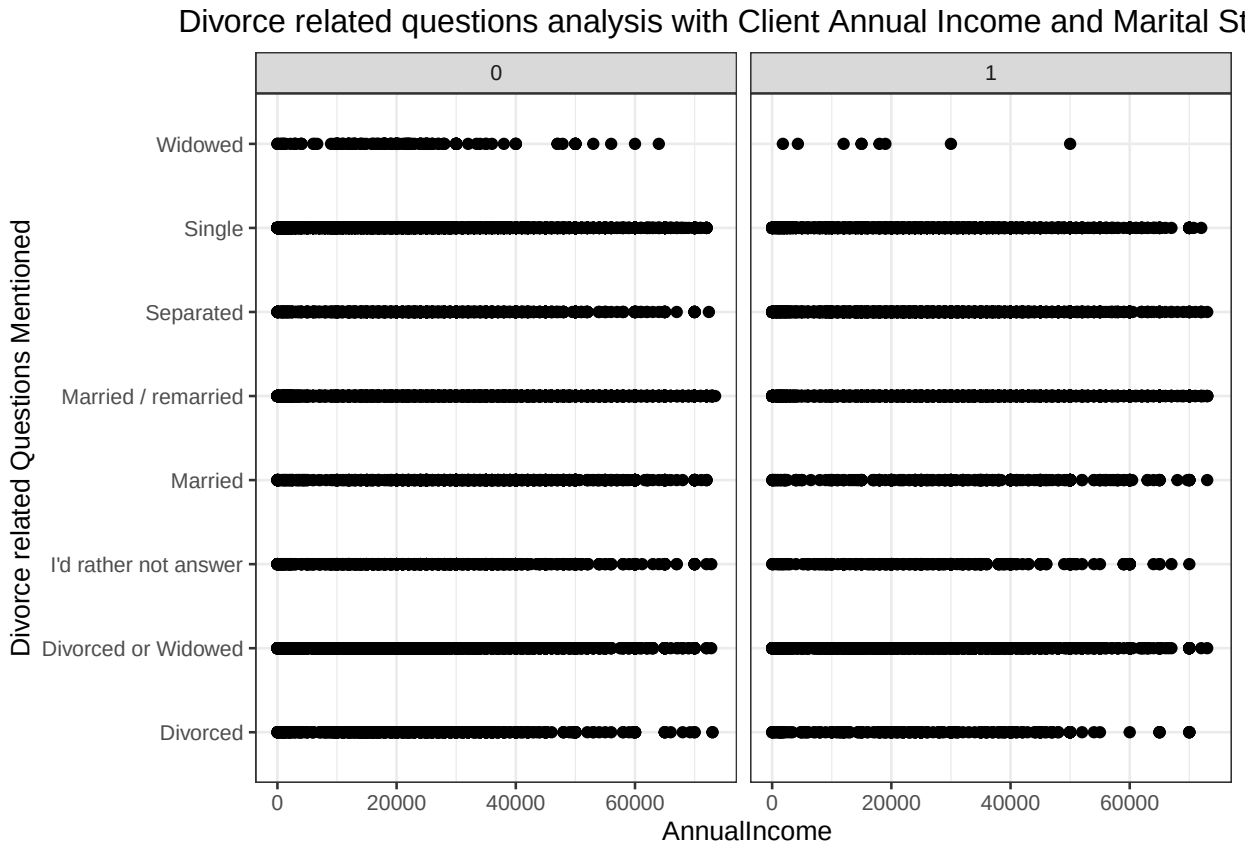
Also the project instructions say to use ggplot2 for our visualization. I find that plot_usmap is its own package so should we try to create a us map plot using ggplot2?

Visualization 3: Financial Status Correlation with Divorce Rate

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

```
# Visualization 3
# 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 & AnnualIncome > 0)

ggplot(data = data_noOutlier, aes(x = AnnualIncome , y = MaritalStatus)) +
  geom_point() +
  labs(x = " AnnualIncome",
       y = "Divorce related Questions Mentioned",
       title = "Divorce related questions analysis with Client Annual Income and Marital Status") +
  facet_wrap(~ Subcategory) +
  theme_bw(base_size = 10) +
  theme(plot.title =
          element_text(hjust = 0.5))
```



Testing models We might need to remove the StateAbbr "US"???

Model 1

```
##
## Call:
## glm(formula = factor(Subcategory) ~ . - Category, family = "binomial",
##     data = data_overall)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0335  -0.8225  -0.4699   0.9026   2.8280
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.268e+00  1.427e-01   8.883  < 2e-16 ***
## StateAbbrAL                    -5.457e-01  1.905e-01  -2.864 0.004178 **
## StateAbbrAR                     2.229e-02  1.618e-01   0.138 0.890438
## StateAbbrAZ                    -9.331e-02  1.298e-01  -0.719 0.472120
## StateAbbrCA                    -1.818e+01  2.397e+02  -0.076 0.939533
## StateAbbrCT                    -3.559e-01  1.724e-01  -2.064 0.038991 *
## StateAbbrFL                    -2.319e-01  1.153e-01  -2.011 0.044329 *
## StateAbbrGA                    -2.881e-01  1.230e-01  -2.342 0.019189 *
## StateAbbrHI                     5.502e-01  1.389e-01   3.961 7.47e-05 ***
## StateAbbrIA                    -1.978e-01  1.460e-01  -1.354 0.175608
## StateAbbrID                    -1.622e+01  6.523e+03  -0.002 0.998015
## StateAbbrIL                     2.909e-01  1.163e-01   2.502 0.012342 *
## StateAbbrIN                    -1.758e+00  1.233e-01 -14.257  < 2e-16 ***
## StateAbbrKS                    -2.708e-01  2.306e-01  -1.174 0.240235
## StateAbbrLA                     1.748e-01  1.332e-01   1.313 0.189295
## StateAbbrMA                    -4.889e-01  1.215e-01  -4.024 5.73e-05 ***
## StateAbbrMD                    -2.588e-01  1.380e-01  -1.875 0.060768 .
## StateAbbrME                     4.665e-01  1.241e-01   3.759 0.000170 ***
## StateAbbrMI                    -3.228e-01  2.178e-01  -1.482 0.138445
## StateAbbrMO                    -3.099e-01  1.206e-01  -2.569 0.010211 *
## StateAbbrMS                     2.666e-01  1.684e-01   1.583 0.113376
## StateAbbrNC                    -4.415e-01  1.207e-01  -3.657 0.000255 ***
## StateAbbrNE                    -1.833e+00  1.525e-01 -12.016  < 2e-16 ***
## StateAbbrNH                    -4.260e-01  1.521e-01  -2.800 0.005115 **
## StateAbbrNJ                    -8.504e-01  2.089e-01  -4.072 4.67e-05 ***
## StateAbbrNM                    -4.418e-02  1.769e-01  -0.250 0.802813
## StateAbbrNY                    -7.817e-01  1.265e-01  -6.178 6.51e-10 ***
## StateAbbrOK                    -5.345e-01  1.327e-01  -4.029 5.60e-05 ***
## StateAbbrPA                    -5.193e-01  1.973e-01  -2.631 0.008505 **
## StateAbbrSC                    -1.851e+01  1.151e+02  -0.161 0.872232
## StateAbbrSD                     4.793e-01  1.673e-01   2.865 0.004174 **
## StateAbbrTN                    -2.159e-01  1.195e-01  -1.807 0.070773 .
## StateAbbrTX                     2.812e-02  1.152e-01   0.244 0.807252
## StateAbbrUS                    -1.832e+01  6.265e+02  -0.029 0.976669
## StateAbbrUT                    -2.468e-01  1.388e-01  -1.778 0.075392 .
## StateAbbrVA                    -3.703e-01  1.238e-01  -2.990 0.002786 **
## StateAbbrVT                    -1.023e-01  1.613e-01  -0.635 0.525735
## StateAbbrWI                    -1.834e+01  1.240e+02  -0.148 0.882402
## StateAbbrWV                    -2.258e-01  1.607e-01  -1.405 0.159895
## StateAbbrWY                     3.285e-01  1.471e-01   2.233 0.025579 *
## Age                            -4.102e-02  9.918e-04 -41.358  < 2e-16 ***
```

```
## NumberInHousehold                   5.483e-02  3.326e-02   1.649 0.099221 .
## MaritalStatusDivorced or Widowed    1.977e-01  8.140e-02   2.429 0.015128 *
## MaritalStatusI'd rather not answer -6.152e-01  1.006e-01  -6.116 9.59e-10 ***
## MaritalStatusMarried               -2.361e-01  9.994e-02  -2.362 0.018164 *
## MaritalStatusMarried / remarried   -5.627e-02  8.057e-02  -0.698 0.484936
## MaritalStatusSeparated              1.559e+00  8.589e-02  18.148  < 2e-16 ***
## MaritalStatusSingle                -1.002e+00  8.047e-02 -12.449  < 2e-16 ***
## MaritalStatusWidowed               -1.577e+00  3.429e-01  -4.598 4.27e-06 ***
## AnnualIncome                       -1.233e-07  5.323e-07  -0.232 0.816766
## SavingsBalance                     -1.845e-05  4.501e-06  -4.099 4.15e-05 ***
## CheckingBalance                    -3.924e-05  8.668e-06  -4.527 5.97e-06 ***
## AllowedIncome                       2.045e-06  2.465e-06   0.829 0.406846
## ---
## 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 50567  degrees of freedom
## AIC: 49539
##
## Number of Fisher Scoring iterations: 17
```

Assumes that there is no severe multicollinearity Assume

Model 2

```
##
## Call:
## glm(formula = factor(Subcategory) ~ . - Category - StateAbbr,
##     family = "binomial", data = data_overall)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2004  -0.8162  -0.6545   1.1211   2.7683
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         5.506e-01  8.636e-02   6.376 1.82e-10 ***
## Age                                -3.522e-02  9.231e-04 -38.152  < 2e-16 ***
## NumberInHousehold                   2.278e-01  1.999e-02  11.398  < 2e-16 ***
## MaritalStatusDivorced or Widowed    2.382e-01  7.609e-02   3.131  0.00174 **
## MaritalStatusI'd rather not answer -5.214e-01  9.487e-02  -5.497 3.87e-08 ***
## MaritalStatusMarried               -2.270e-01  9.338e-02  -2.431  0.01507 *
## MaritalStatusMarried / remarried   -1.025e-02  7.525e-02  -0.136  0.89161
## MaritalStatusSeparated              1.311e+00  7.887e-02  16.618  < 2e-16 ***
## MaritalStatusSingle                -8.750e-01  7.529e-02 -11.622  < 2e-16 ***
## MaritalStatusWidowed               -1.531e+00  3.364e-01  -4.552 5.32e-06 ***
## AnnualIncome                       -7.687e-07  5.054e-07  -1.521  0.12829
## SavingsBalance                     -1.278e-05  4.201e-06  -3.043  0.00234 **
## CheckingBalance                    -3.408e-05  8.221e-06  -4.146 3.38e-05 ***
## AllowedIncome                      -1.134e-05  1.389e-06  -8.160 3.34e-16 ***
```

```
## ---
## 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: 56542  on 50606  degrees of freedom
## AIC: 56570
##
## Number of Fisher Scoring iterations: 5
```