General goal: The ABA would like to anticipate the sorts of legal questions that arise so that they can prepare volunteers to address those questions, better know how and when to recruit lawyers with specific expertise, and know how to advise state partners on general trends they’re seeing.

Specific Listed Goals:

* Provide a better way for attorneys to better connect with their clients
  + Match similar demographic attorneys to their clients?
    - Based on race, historical background
  + Provide a listed background synopsis of the client to better
* Information on ABA Video
  + Largest professional membership for lawyers
  + Provide pro bono services to low income people
  + Encourage lawyers to provide service to them
  + Lets people have legal questions and volunteer attorney will answer
  + Provides help to those who cannot get in-person services
  + Client population, demographics, and personal information on why and how they came across the service would better train and recruit attorneys
  + Goals:
    - Want to reach out to people who need it
    - Provide a way to balance attorney resources
    - Advertise better to the right demographic
    - LIMITED RESOURCES, want to better provide services
* **DATA IMPORTANT NOTES:**
  + Closed conversation
    - Requested by client, attorney, or admin **OR** no communication for 10 days
* **Plan**
  + analyze/clean data sets
    - Possible Questions:
      * Is there a correlation between number of clients and attorneys in a state
      * Correlation between clients and attorneys in each county?
      * Correlation between clients and attorneys in postal code?
      * Correlation between states’ financial well-being and client OR attorney
      * Correlation between question amount and client proportion to the state population?
      * Correlation between question amount and state financial well-being
      * Correlation between gender and question type
      * Correlation between number of unsolved interactions and some other factor?
    - 1: Attorneys, AttorneyTimeEntries (Luke)
      * Attorneys
        + ID is indexing of each observation
        + 11544 attorneys
        + First attorney joined program: 2011-04-08
        + Last attorney joined program: 2022-01-24
        + 42 States only included for attorneys (40 *actual* states, 1 territory, and one federal)

Used state.abb as basis for state data

states that aren’t considered as a “state”

VI, US

Virgin islands and US in general

States with no attorneys

"CO" "DE" "KY" "MN" "MT" "NV" "OH" "OR" "RI" "WA"

* + - * + 742 unique counties
        + 2275 unique cities
        + NO NAs in any column
        + Chicago has the most attorneys
      * AttorneyTimeEntries
        + Texas has the most logged hours

**Attorney Sum per State**

state sum

1 TX 1140

2 FL 1103

3 TN 1036

4 IL 997

5 NC 629

6 MA 533

7 VA 517

8 IN 484

9 NY 476

10 WI 385

11 MO 315

12 GA 308

13 NE 266

14 MD 262

15 SC 260

16 LA 246

17 CA 245

18 ME 193

19 AR 192

20 WV 175

21 OK 167

22 AL 138

23 MS 138

24 HI 125

25 NH 123

26 AK 110

27 AZ 104

28 CT 103

29 PA 98

30 UT 97

31 NM 94

32 MI 77

33 NJ 74

34 IA 67

35 WY 62

36 VT 54

37 KS 30

38 SD 26

39 VI 3

40 ID 2

41 ND 1

* + - 2: Categories, Subcategories (Zhou)
    - 3: Questions, QuestionPosts (Shawn)
      * QuestionPosts: 505564 entities, but only 164735 QuestionUno’s
    - 4: Clients (Kevin)
    - 5: StateSites (Tian)

50 states: "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD", "MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ", "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV", "WI", "WY"

Weird Client: 90D7EB25-FAAE-415E-9878-FB8655EE672F

**Clients Who Asked Questions**

AnnualIncome Number of Clients Proportion of Clients

Below 5k 19343 0.1168

5k-10k 16625 0.1004

10k-20k 44819 0.2707

20k-30k 40277 0.2433

30k-40k 22769 0.1375

40k-50k 12093 0.0730

50k-100k 9357 0.0565

100k-300k 254 0.0015

Above 300k 16 0.0001

**Clients Who Had Their Question Answered**

AnnualIncome Number of Clients Proportion of Clients

Below 5k 13444 0.1115

5k-10k 11781 0.0977

10k-20k 32324 0.2681

20k-30k 29659 0.2460

30k-40k 17004 0.1411

40k-50k 9099 0.0755

50k-100k 7034 0.0584

100k-300k 186 0.0015

Above 300k 14 0.0001

**Total Clients**

AnnualIncome Number of Clients Proportion of Clients

Below 5k 27929 0.1118

5k-10k 22133 0.0886

10k-20k 57969 0.2320

20k-30k 52239 0.2091

30k-40k 33191 0.1328

40k-50k 21271 0.0851

50k-100k 29727 0.1190

100k-300k 4705 0.0188

Above 300k 716 0.0029

**Mean Hours Worked By Attorney Per State**

Hours

StateAbbr

SD 92.02

HI 22.74

NJ 20.07

MO 19.69

TX 17.70

CA 15.41

WI 14.51

NM 14.44

VA 13.45

SC 13.02

FL 12.30

NY 12.22

MA 11.74

LA 11.72

OK 10.69

WV 10.47

IN 10.42

ME 10.18

CT 9.83

VT 9.43

NC 8.77

TN 8.75

IL 8.74

AZ 8.04

WY 7.40

NE 6.80

IA 6.71

UT 6.41

MI 6.18

AR 6.03

NH 5.79

MD 5.77

AK 5.27

GA 4.88

PA 4.61

AL 3.79

MS 3.54

US 3.00

KS 0.83

**Mean Hours Worked Per Attorney Per Session By State**

Hours

StateAbbr

MI 1.3600

CA 1.3011

SD 0.9895

TX 0.6214

MA 0.5385

US 0.5376

PA 0.4987

VA 0.4935

MS 0.4735

NE 0.4293

IL 0.4190

MD 0.4112

IN 0.4054

AZ 0.4027

WY 0.3933

GA 0.3909

CT 0.3731

ME 0.3713

KS 0.3667

SC 0.3635

TN 0.3572

HI 0.3541

NJ 0.3534

LA 0.3439

WV 0.3380

NH 0.3305

AK 0.3222

NY 0.3170

VT 0.3145

WI 0.3094

NC 0.3086

OK 0.2960

AL 0.2871

NM 0.2805

FL 0.2765

MO 0.2758

UT 0.2724

AR 0.2642

IA 0.2392

**Total Hours Worked By All Attorneys By State**

Hours

StateAbbr

TX 5998.70

FL 3689.84

IL 3680.30

TN 2826.86

VA 2286.60

IN 2228.92

MO 2145.80

MA 2043.25

WI 1929.20

NC 1816.05

NY 1760.10

SD 1748.40

SC 1705.97

CA 1310.20

LA 1077.85

HI 886.90

ME 844.80

WV 733.10

MD 611.50

NE 604.90

OK 577.00

AR 530.50

GA 497.60

NM 433.10

CT 402.90

NH 364.50

NJ 361.20

WY 281.20

AZ 265.40

VT 217.00

AK 184.60

IA 167.70

MS 166.20

AL 163.10

UT 141.10

PA 78.30

MI 68.00

US 62.90

KS 6.60

**All Clients**

Gender Number of Clients Proportion of Clients

Female 185433 0.66

Male 87517 0.31

I'd rather not answer 6257 0.02

Other 796 0.00

**Individual Households**

Gender Number of Clients Proportion of Clients

Female 39963 0.58

Male 26426 0.39

I'd rather not answer 1779 0.03

Other 282 0.00

**Family Households**

Gender Number of Clients Proportion of Clients`

Female 130929 0.70

Male 53058 0.28

I'd rather not answer 3526 0.02

Other 456 0.00

**MUST SHOW THAT FAMILY HOUSEHOLDS HAVE HIGHER FEMALE PROPORTION OF CLIENTS/FEMALES MORE LIKELY TO REPRESENT THEIR HOUSEHOLD/PROPORTION OF FEMALES IN FAMILY HOUSEHOLD IS GREATER THAN FEMALE PROPORTION OF INDIVIDUAL CLIENTS**

**Null (No answer) Households**

Gender Number of Clients Proportion of Clients

NaN 27224 0.54

Female 14541 0.29

Male 8033 0.16

I'd rather not answer 952 0.02

Other 58 0.00

**Gender Data For Marriage Status**

Female Male

# of Married Clients 47239 28466

Married Clients Proportions 0.62 0.37

# of Single Clients 75783 37791

Single Clients Proportions 0.67 0.33

# of Divorced/Separated Clients 55438 18286

Divorced/Separated Clients Proportions 0.75 0.25

**Female Most Common Question Categories**

Number of Clients Proportion of Clients

Category

Family and Children 47959 0.47

Housing and Homelessness 18159 0.18

Other 16559 0.16

Consumer Financial Questions 7907 0.08

Work, Employment and Unemployment 4644 0.05

Individual Rights 3464 0.03

Health and Disability 1254 0.01

Income Maintenance 1179 0.01

Education 411 0.00

Juvenile 167 0.00

**Male Most Common Question Categories**

Number of Clients Proportion of Clients

Category

Family and Children 15821 0.36

Other 9371 0.21

Housing and Homelessness 7429 0.17

Consumer Financial Questions 4547 0.10

Work, Employment and Unemployment 2970 0.07

Individual Rights 2357 0.05

Income Maintenance 680 0.02

Health and Disability 616 0.01

Education 135 0.00

Juvenile 55 0.00

**No Gender Given Most Common Question Categories**

Number of Clients Proportion of Clients

Category

Family and Children 12465 0.47

Other 4711 0.18

Housing and Homelessness 4379 0.16

Consumer Financial Questions 2362 0.09

Work, Employment and Unemployment 1157 0.04

Individual Rights 827 0.03

Health and Disability 346 0.01

Income Maintenance 316 0.01

Education 136 0.01

Juvenile 74 0.00

**Gender Proportion Among All Clients**

Gender Number of Clients Proportion of Clients

0 Female 185433 0.6617

1 Male 87517 0.3123

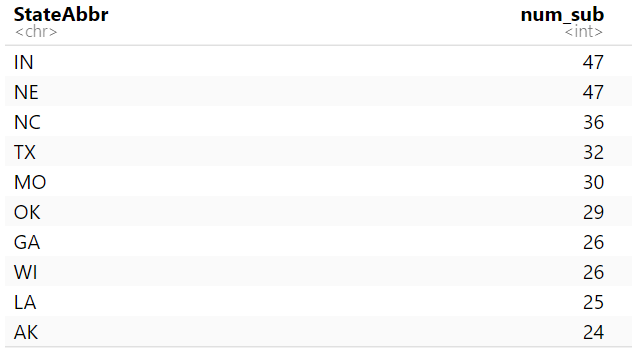
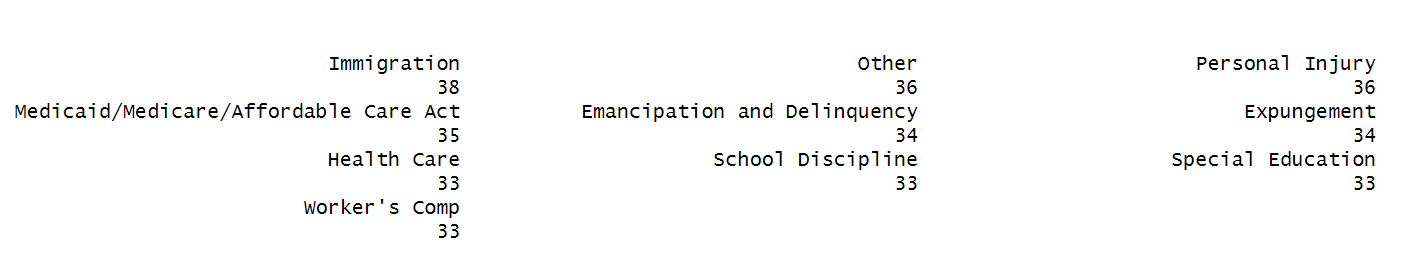
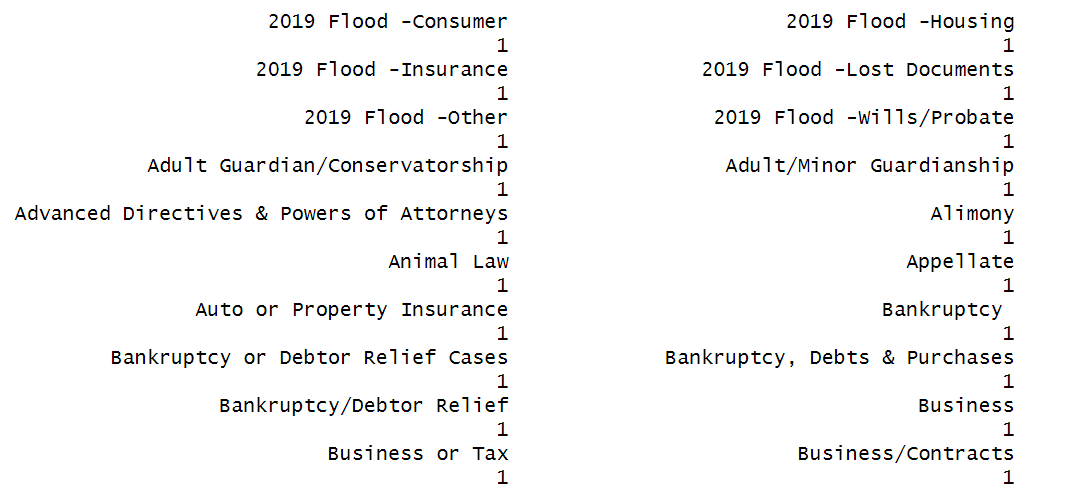
2 Nonconforming 7277 0.0260

# Categories & Subcategories

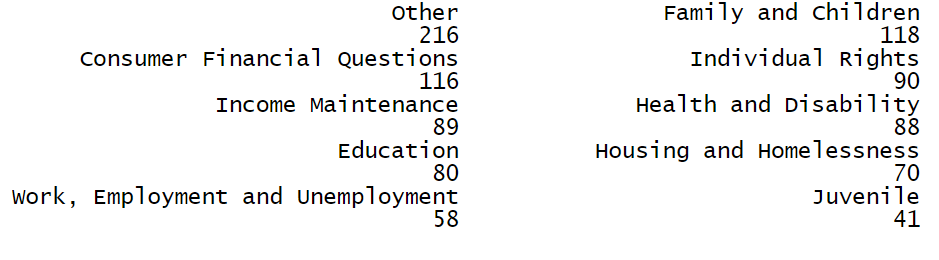
## Categories

* There are 41 states and 2 regions (US, VI)
  + Missing states:
    - CO: Colorado
    - DE: Delaware
    - KY: Kentucky
    - MN: Minnesota
    - MT: Montana
    - NV: Nevada
    - OH: Ohio
    - OR: Oregon
    - RI: Rhode Island
* There are 10 categories in each state/region:
  + Consumer Financial Questions
  + Education
  + Work
  + Employment and Unemployment
  + Family and Children
  + Health and Disability
  + Juvenile
  + Housing and Homelessness
  + Income Maintenance
  + Individual Rights
  + Other

## Subcategories

* There are 365 unique subcategories
* Number of subcategories in each state
  + Top 10:
  + 
  + Bottom 10
  + 
* Frequencies
  + Top 10 most frequently subcategories
  + 
  + Least frequently
  + 

## Both

* Number of subcategories in each category
  + 

**Potential Directions**

* Most frequently asked categories/subcategories
  + Relationship between the category of questions and financial status of individuals?
* No. of subcategories
  + (under the same category), correlation between no. of subcategories, no. of attorney, and no. of questions asked
* Gender ratio in clients who asked questions
  + Is gender correlated to…
    - Specific categories of questions?
    - Financial status?
    - Tone of questions asked?
* Unanswered questions
  + Related to the attorney number in that state?
  + Concentrated in specific states?
  + Concentrated in specific categories?
* Where does the silent majority go?
* Correlation of total logged hours per state vs. client amount
* Sentiment analysis
  + Correlation between tones and categories/subcategories
* 3 Types of Interactions
  + Clients who do not ask
  + Clients who ask but do not receive response
  + Clients who have at least interacted with
* Time related
  + Do gender

**Questions to solve**

* Attorney
  + Recruit more
  + Allocation to their specialized field
* Client
  + Convert more silent clients into asking questions
* Service
  + Reduce more unsolved conversations
* Trends
  + General patterns observed

**Plan**

* Why is there a gender imbalance in the client amount?
  + Does gender have a correlation to the type of questions being asked?
  + External factors to why there is a skew in amount of woman clients and men clients
* Places that need special attention:
  + Rank the subcategory difficulties (categories)
    - Convo length, question number, emotional intensity
  + Sentiments, time-spans, frequencies of each sub-category

-> property of each sub-category

-> suggestions on human resources allocation and how to connect to the clients better

**Rubric**

Ratio

> # p checking: more female proportion clients in family households

> # vs individual households

> prop.test(x = c(39963, 130929), n = c(68450, 187969), alternative = "l")

2-sample test for equality of proportions with continuity

correction

data: c(39963, 130929) out of c(68450, 187969)

X-squared = 2867.4, df = 1, p-value < 2.2e-16

alternative hypothesis: less

95 percent confidence interval:

-1.000000 -0.109152

sample estimates:

prop 1 prop 2

0.5838276 0.6965457

> # p checking: more female prop clients in divorced/sep than married/single

> # female divorced/sep: 55438, total divorced/sep: 73724

> # female married/single: 75783, total married/single: 113574

> prop.test(x = c(55438,75783), n = c(73724, 113574), alternative = "g")

2-sample test for equality of proportions with continuity

correction

data: c(55438, 75783) out of c(73724, 113574)

X-squared = 1528.9, df = 1, p-value < 2.2e-16

alternative hypothesis: greater

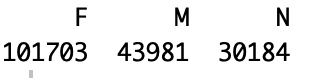
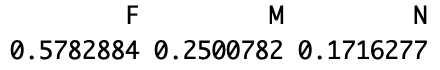
95 percent confidence interval:

0.08121567 1.00000000

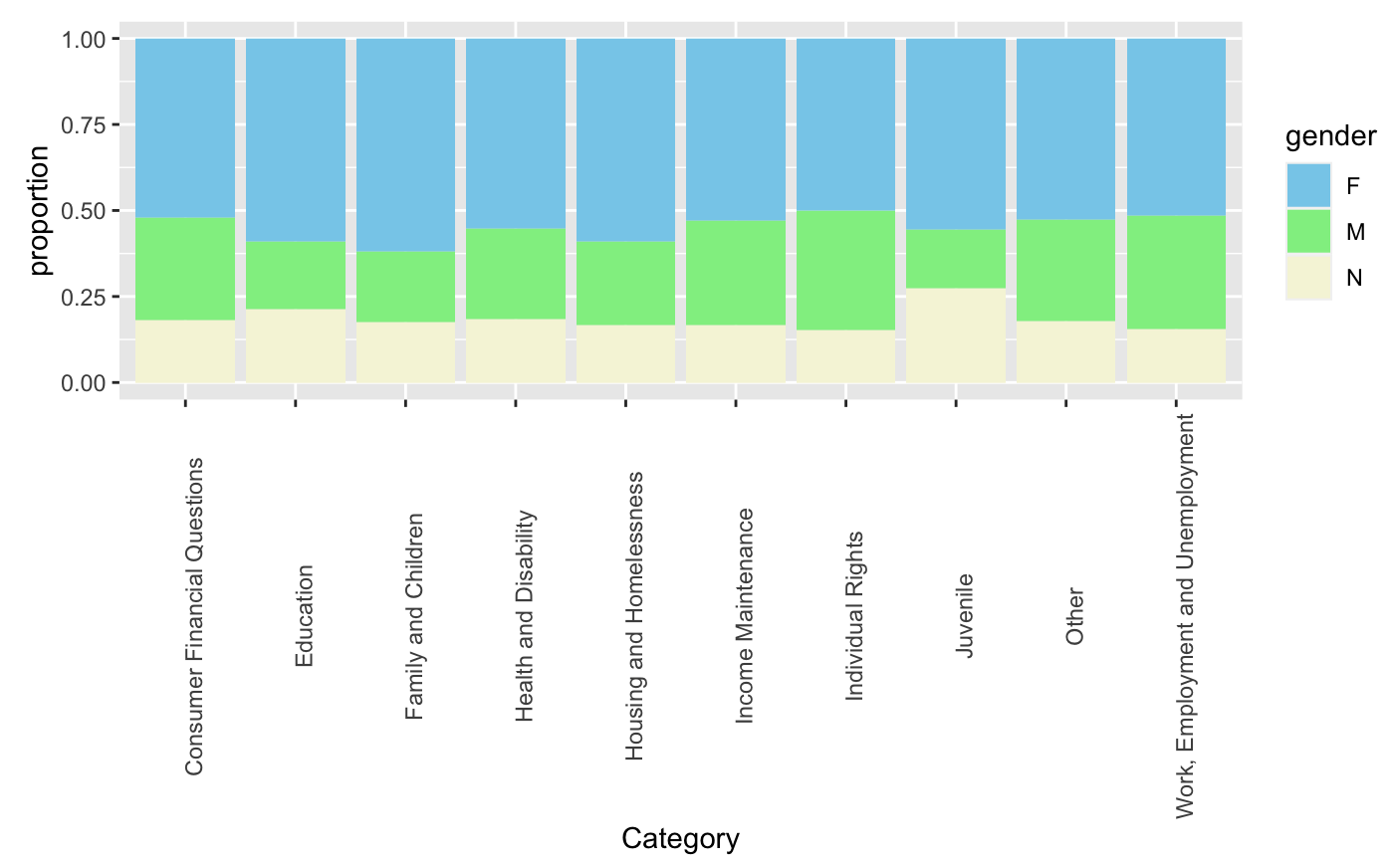
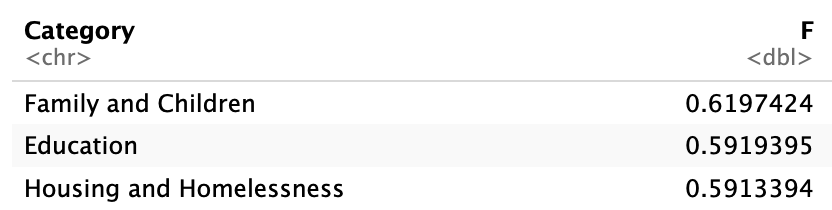
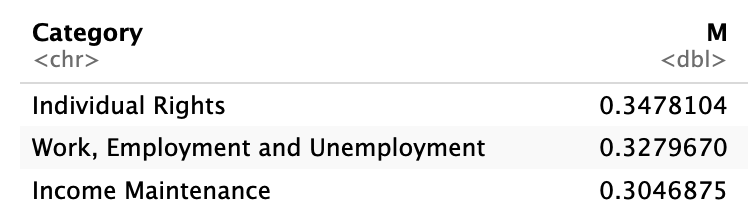
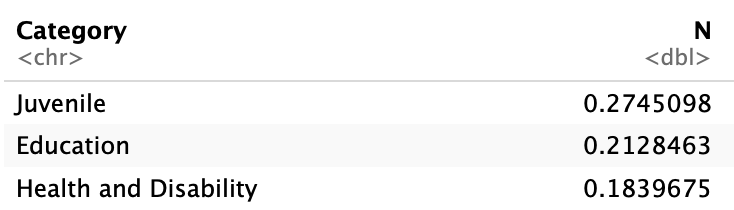
sample estimates:

prop 1 prop 2

0.7519668 0.6672566

* More female clients
  + 57.8% female clients
  + 2.3-to-1 female-male ratio
    - 
    - 
    - Potential explanations…
      * Women face more legal issues
      * Women more likely to seek legal assistance
        + Gender role and societal expectation

Households - higher proportion of female clients

* + Genders - categories of questions
    - 
    - Top three categories females ask about
      * 
    - Top three categories males ask about
      * 
    - Top three categories non-conforming genders ask about
      * 
* ‘difficulty index’ for categories
  + Sentiment
    - Must test the general idea that categories’ means reflect their true sentimental mean

> greaterTTest(juv\_sent$sentiment, fam\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 1.3992, df = 69820, p-value = 0.08088

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.009916489 Inf

sample estimates:

mean of x mean of y

0.4767442 0.4202749

Welch Two Sample t-test

data: v1 and v2

t = 1.3343, df = 258.74, p-value = 0.09164

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.01339293 Inf

sample estimates:

mean of x mean of y

0.4767442 0.4202749

> ## special case needed to check

> greaterTTest(juv\_sent$sentiment, other\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 2.1469, df = 31131, p-value = 0.0159

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.0198469 Inf

sample estimates:

mean of x mean of y

0.4767442 0.3918704

Welch Two Sample t-test

data: v1 and v2

t = 2.0016, df = 260.74, p-value = 0.02318

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.01487862 Inf

sample estimates:

mean of x mean of y

0.4767442 0.3918704

> greaterTTest(fam\_sent$sentiment,other\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 6.4661, df = 100437, p-value = 5.051e-11

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.02117881 Inf

sample estimates:

mean of x mean of y

0.4202749 0.3918704

Welch Two Sample t-test

data: v1 and v2

t = 6.5248, df = 60470, p-value = 3.432e-11

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.02124375 Inf

sample estimates:

mean of x mean of y

0.4202749 0.3918704

> greaterTTest(other\_sent$sentiment,individual\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 1.9942, df = 37994, p-value = 0.02307

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.002890737 Inf

sample estimates:

mean of x mean of y

0.3918704 0.3753686

Welch Two Sample t-test

data: v1 and v2

t = 2.0212, df = 10817, p-value = 0.02164

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.003071795 Inf

sample estimates:

mean of x mean of y

0.3918704 0.3753686

> # HIGH P VALUE

> greaterTTest(individual\_sent$sentiment,income\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 0.12115, df = 9389, p-value = 0.4518

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.02264506 Inf

sample estimates:

mean of x mean of y

0.3753686 0.3735683

Welch Two Sample t-test

data: v1 and v2

t = 0.12199, df = 3865.7, p-value = 0.4515

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.02248132 Inf

sample estimates:

mean of x mean of y

0.3753686 0.3735683

> greaterTTest(income\_sent$sentiment, consumer\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 4.62, df = 15997, p-value = 1.933e-06

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.03948041 Inf

sample estimates:

mean of x mean of y

0.3735683 0.3122587

Welch Two Sample t-test

data: v1 and v2

t = 4.4628, df = 2990.1, p-value = 4.194e-06

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.03870567 Inf

sample estimates:

mean of x mean of y

0.3735683 0.3122587

> # HIGH P VALUE

> greaterTTest(consumer\_sent$sentiment, health\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 0.15185, df = 16025, p-value = 0.4397

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.01955719 Inf

sample estimates:

mean of x mean of y

0.3122587 0.3102698

Welch Two Sample t-test

data: v1 and v2

t = 0.15234, df = 3123.6, p-value = 0.4395

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.01949296 Inf

sample estimates:

mean of x mean of y

0.3122587 0.3102698

> greaterTTest(health\_sent$sentiment, work\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 3.3957, df = 11026, p-value = 0.0003435

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.02271645 Inf

sample estimates:

mean of x mean of y

0.3102698 0.2662085

Welch Two Sample t-test

data: v1 and v2

t = 3.2837, df = 3451.3, p-value = 0.0005174

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.02198446 Inf

sample estimates:

mean of x mean of y

0.3102698 0.2662085

> greaterTTest(work\_sent$sentiment, house\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 7.4707, df = 37971, p-value = 4.074e-14

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.03681992 Inf

sample estimates:

mean of x mean of y

0.2662085 0.2189926

Welch Two Sample t-test

data: v1 and v2

t = 7.1928, df = 13572, p-value = 3.34e-13

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.03641779 Inf

sample estimates:

mean of x mean of y

0.2662085 0.2189926

> # HIGH P VALUE

> greaterTTest(house\_sent$sentiment,edu\_sent$sentiment)

Two Sample t-test

data: v1 and v2

t = 0.13848, df = 29861, p-value = 0.4449

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.03115161 Inf

sample estimates:

mean of x mean of y

0.2189926 0.2161290

Welch Two Sample t-test

data: v1 and v2

t = 0.1375, df = 645.13, p-value = 0.4453

alternative hypothesis: true difference in means is greater than 0

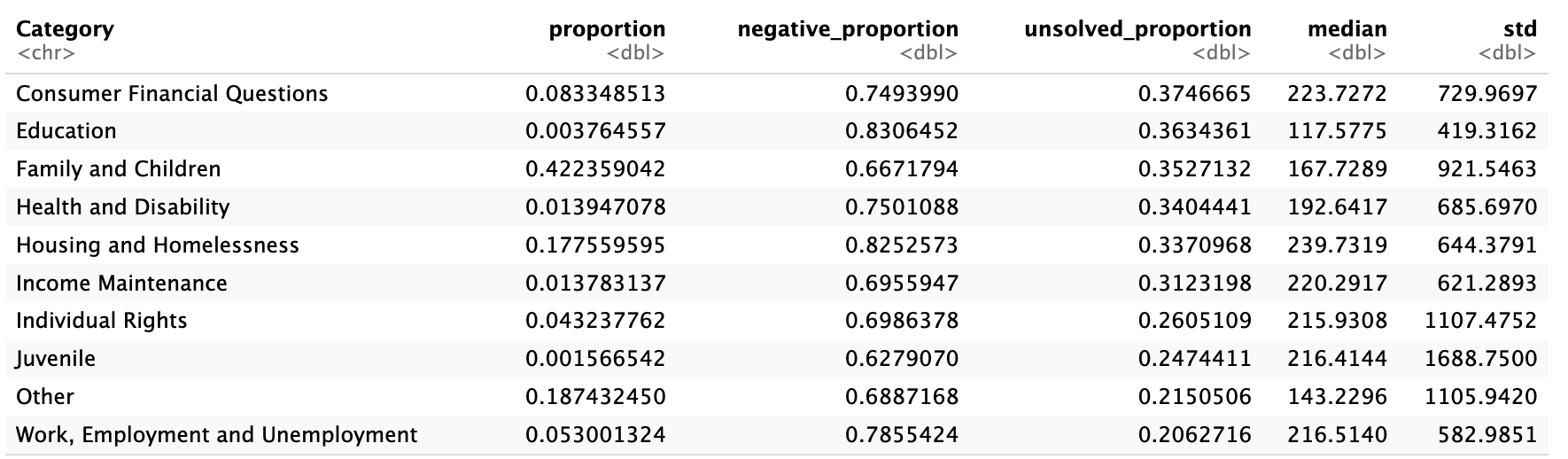
95 percent confidence interval:

-0.03144051 Inf

sample estimates:

mean of x mean of y

0.2189926 0.2161290

* + Question quantity (tied to states?)
  + Quantity of unsolved cases
  + Timespan median
  + Timespan sd
  + 

Gender:

* Barplots
  + Female proportion vs. Marriage status
  + Female proportion in single vs. Female proportion in household
  + Female proportion in Categories
  + Female proportion in total client

Difficulties:

Consumer Financial Questions

Education

Work

Employment and Unemployment

Family and Children

Health and Disability

Juvenile

Housing and Homelessness

Income Maintenance

Individual Rights

Other

* Slide1
  + Initial thoughts on the data
    - Noticed gender imbalanced gender proportion in clients
    - Looked into other variables to see if other factors changed this claim
  + Observations on other factors based on gender within specific samples
    - Nonconforming consists of responses that were not Female, Male, or Null
    - All Null responses were excluded (clients who skipped the question without responding)
    - Initially, we looked to see the total client gender proportion. The pie chart provides the observed proportion.
    - The following variables were used to see the gender distribution, and see if the factors have a significant effect on gender proportion
    - LSC Categories
      * We observed differences in the proportion of female clients within each category. While we can tentatively put forth possible reasons for the difference such as a higher rate of women seeking divorce than men, there are too many possible confounding variables to confidently state a cause for the imbalanced ratio between genders.
    - Marital Status
      * There was a difference in female/non-female proportions based on marital status. Looking at the observed proportions, we decided to group together married & single, and divorced & separated for they had similar proportions. We executed a proportion test and the result implies that there is a higher probability to get a female client in the divorced/separated marital status rather than married/single.
    - Household
      * In addition, there was an observable difference in household status for the gender proportion. The proportion test concluded that there is a higher chance to get a female client in a family household than an individual household.
      * These two client demographics are some of the major factors that seem to have a direct effect on the gender distribution of clients.
* Slide2
  + We now switch gear to examine if it is harder for attorneys to answer questions from specific categories. As you may see in this 5-dimensional bubble chart, we gathered five variables as potential indicators of difficulty level, starting with total\_proportion on the horizontal axis, which is the proportion of total questions asked under each category. **(17 seconds)**
    - Negative proportion - sentiment analysis **(30 seconds)**
      * **(Shown along one axis)**
      * We extracted the negative proportion feature with natural language processing techniques. In particular, we loaded a deberta-v3-large model fine-tuned on sentiment analysis dataset. Then we let the model to predict the sentiment of the clients’ questions and obtain the proportion of negative posts. Questions labeled as negative often contains words that are emotionally strong or describing a severe scenario.
    - Unsolved proportion
      * **(Shown along one axis)**
      * Percentage of questions that were unanswered.
    - Median/STD of time span
      * Finally, we have Median and STD of the time spans of each case represented by the **size and color of the spheres** respectively
  + Observations **(50 seconds)**
    - Family and Children: significantly more questions asked under this category, 23% higher than the second highest (Other). Highest unsolved questions as well (37%). But it does have the lowest negative\_proportion.
      * Potential response: allocate more attorneys or attorneys with specialized skills under this category.
    - Housing and Homelessness: highest time span of each question (240 hours).
      * Possibly by nature of this category and how it has the highest negative proportion, it takes longer time to resolve.
    - Juvenile: extreme variable on the other end. lowest in three axes. However, it does have the highest standard deviation of the time\_span.
  + After ranking the sentimental values of the categories and testing the true difference between expected values, some p-values imply that some categories aren’t statistically significantly different enough while some categories do.
    - This mixed result implies that there is some correlation between the type of category and the expected value of the sentiment value, and so this shows a high likelihood for a possible categorical algorithm that can aid with allocating specific questions to certain attorneys. **(26 seconds)**
  + We can use these features in two levels of granularity. Each separate feature could be used to facilitate a more efficient and empathetic client-attorney relationship, whereas examining all features together can help us identify more challenging categories with larger and darker spheres that are far from the origin.​
  + if ABA would like to adjust its recruitment or resource allocation strategies, it could be valuable to pay more attention to
* Future work
  + Do the same feature extraction for sub-categories
* <https://drive.google.com/drive/folders/13Ecwt4DixkKIU8qNs_X9aO4xD6zcy6Sv>