Stat 302 Project 1

Olivia Fang, Jason Wu and Ziyi Zhao

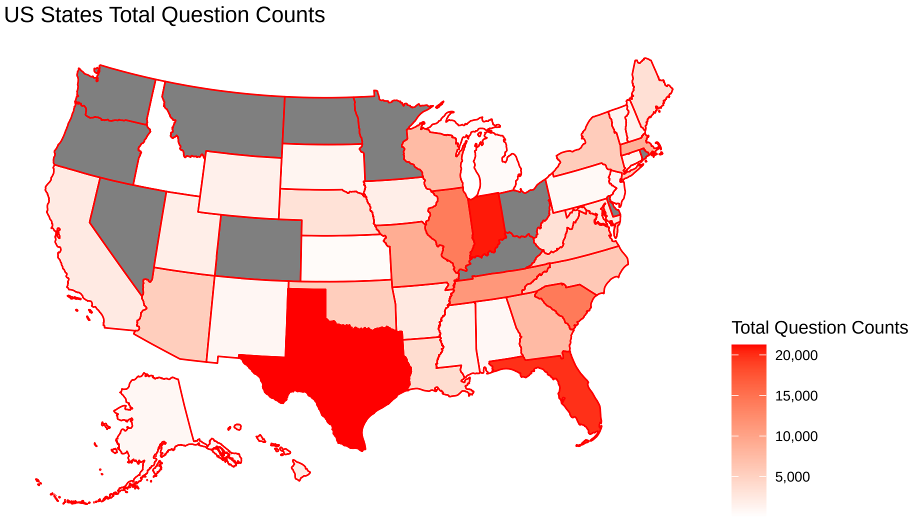
2023-04-29

```
# Please set Working Directory to Source File
attorneys <- read_csv("data/attorneys.csv", show_col_types = FALSE)
attorneytimeentries <- read_csv("data/attorneytimeentries.csv", show_col_types = FALSE)
categories <- read_csv("data/categories.csv", show_col_types = FALSE)
questions <- read_csv("data/questions.csv", show_col_types = FALSE)
```

Data Manipulation and Visualizations

We are interested in the category variables and wonder which category have more questions and conversations between clients and lawyers. Therefore, we first plot a map to examine whether the inquiries would vary by states.

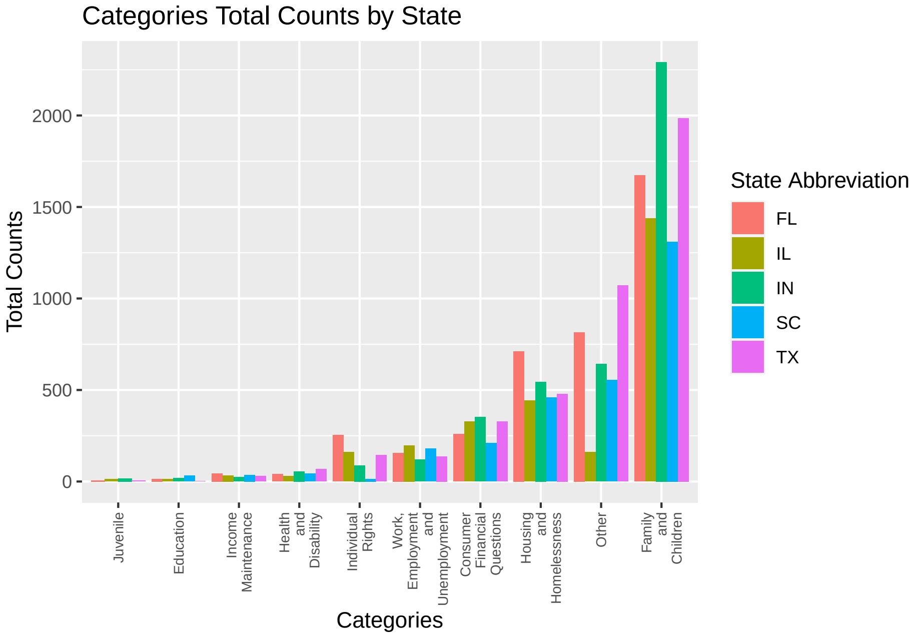
```
# Category Total Count
df_tmp <- questions %>% group_by(StateAbbr) %>% mutate(total_questionCounts = n(),
  state=StateAbbr) %>% distinct(StateAbbr, .keep_all = TRUE)
plot_usmap(data=df_tmp, values = "total_questionCounts", color = "red") +
  labs(title = "US States Total Question Counts") +
  scale_fill_continuous(low = "white", high = "red", name = "Total Question Counts",
  label = scales::comma) + theme(legend.position = "right")
```



After examining the data, we saw importance in the count of questions being asked by various states. Despite the fact that there are some missing values for several state (states with grey area), We could observe that Florida, Illinois, Indiana, Southern Carolina, and Texas have the most significant value of total inquires within the state.

But visualizing the count of questions in the map doesn't tell us much about the other variables. We wanted to see how this ranking would vary by states. Since there are a large number of states, and plotting them all would give no useful information, we focus on the top 5 states that have the most question inquiries.

```r
# Category Total Count Ranked by State
questions_top5 <- questions %>% group_by(StateAbbr) %>%   mutate(total_StateAbbrCounts = n()) %>%
  distinct(total_StateAbbrCounts, .keep_all = TRUE) %>%
  arrange(-total_StateAbbrCounts) %>% pull(StateAbbr) %>% head(5)
subset(questions, questions$StateAbbr == questions_top5) %>%
  ggplot(aes(x = reorder(Category, Category, length), fill = StateAbbr)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(size=7, angle=90, hjust=1.0, vjust=0.5)) +
  labs(title = "Categories Total Counts by State", x = "Categories", y = "Total Counts") +
  scale_x_discrete(labels = wrap_format(5)) + scale_fill_discrete(name = "State Abbreviation")
```

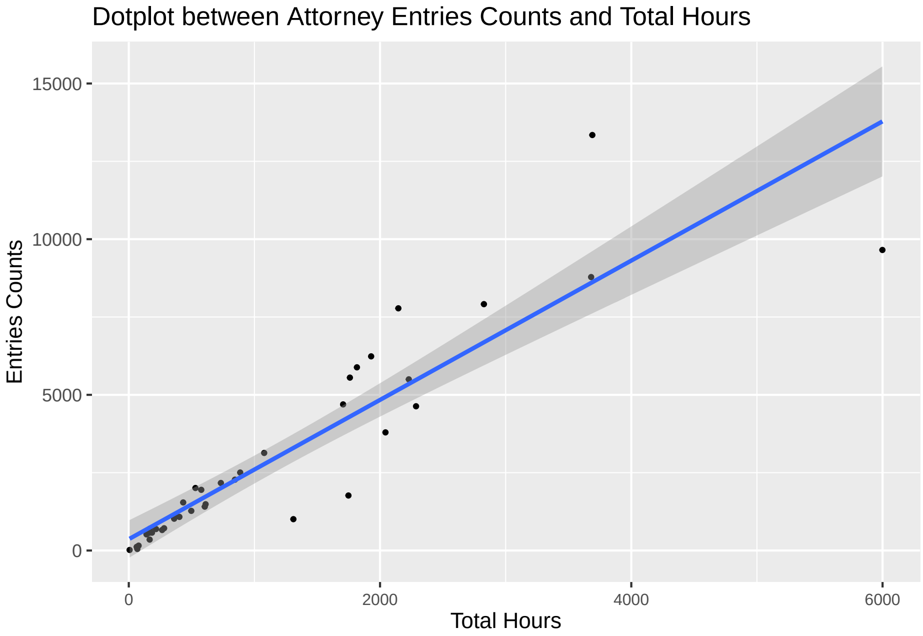


This graph shows a breakdown of all the question categories by the top 5 states (Florida, Illinois, Indiana, Southern Carolina, Texas). The bars are stacked next to each other to show that they belong to the same category, and are color-coded by state so viewers can differ each bar.

We can see that there are no fixed patterns between the each question category and the states. For all five states, the most asked questions is in the Family and Children category. Then for the second most asked category, other, it is also the second most asked category for all the states except Illinois. And for the 1st most asked category, Indiana asked the most questions, while for the 2nd most asked category, Texas asked the greatest number of questions. So there are definitely some correlations between states and category, but not a fixed pattern.

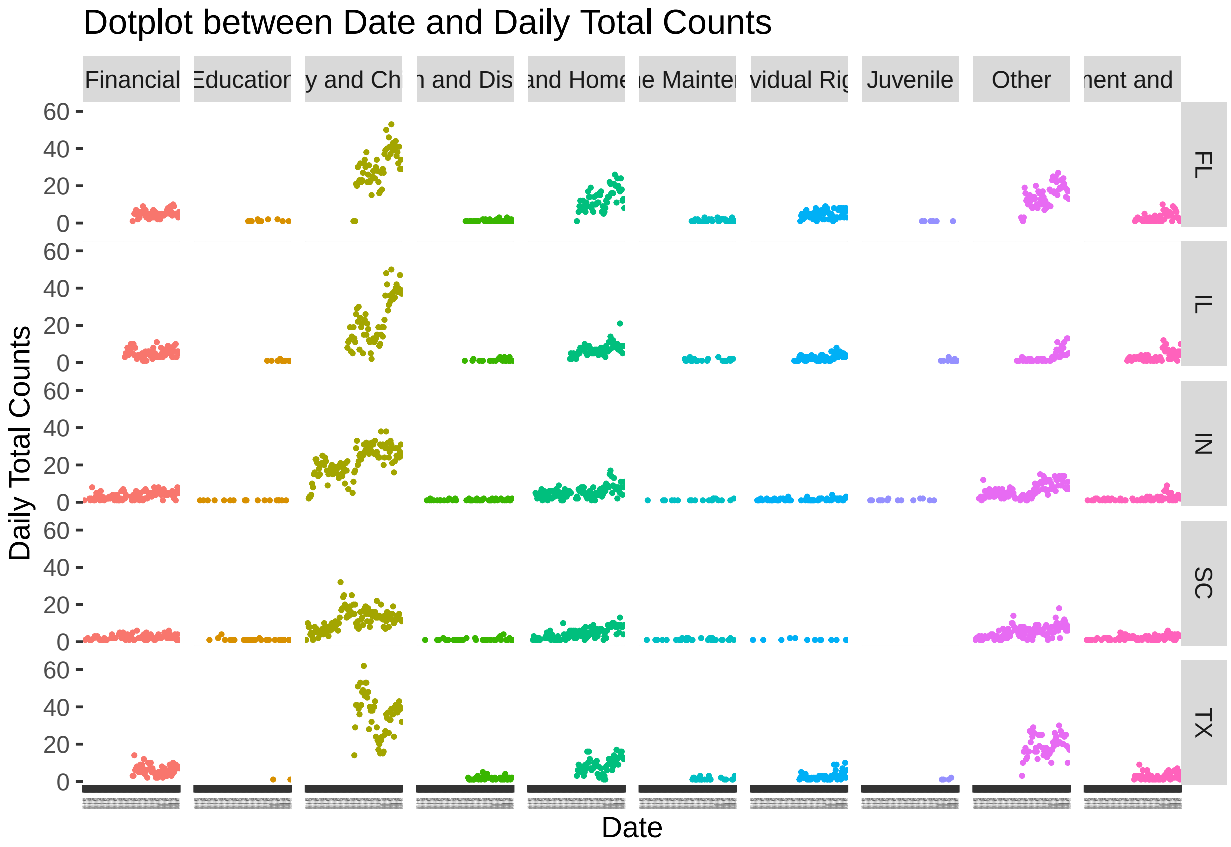
After looking at these visualizations of categories and question counts, we also wondered whether there is correlation between total hours of conversations in each state, and the entries count in each state.

```r
# Relationship between Attorney Entries Counts and Total Hours
AttorneyEntries <- attorneytimeentries %>%
  group_by(StateAbbr) %>%
  mutate(entries_count = n(), total_hours = sum(Hours)) %>%
  distinct(StateAbbr, .keep_all = TRUE) %>%
  select(StateAbbr, entries_count, total_hours)
AttorneyEntries %>%
  ggplot(aes(x = total_hours, y = entries_count)) +
  geom_point(size = 0.8) +
  theme(axis.text.x = element_text(size=8, angle=0, hjust=0.5, vjust=0.5),
        legend.position = "none") +
  labs(title = "Dotplot between Attorney Entries Counts and Total Hours",
       x = "Total Hours", y = "Entries Counts") +
  geom_smooth(formula = 'y ~ x', method = lm)
```



Because we are also trying to find whether total hours can be used as either an independent or a dependent variable to total counts. From the graph below, we found that there's a strong positive relationship between total counts and total hours. That is to say, as Total Hour increase, Total counts increases. We are also eligible to check that there are no cases of entries that has few hours but large total counts.

```
# Relationship between Date and Daily Total Counts
Questions <- (questions)
Questions$AskedOnUtc <- format(as.Date(questions$AskedOnUtc), "%Y-%m")
subset(Questions, Questions$StateAbbr == questions_top5) %>%
  group_by(StateAbbr, AskedOnUtc, Category) %>% mutate(daily_totalCounts = n()) %>%
  distinct(AskedOnUtc, .keep_all = TRUE) %>%
  ggplot(aes(x = AskedOnUtc, y = daily_totalCounts, color = Category)) +
  geom_point(size=0.4) +
  theme(axis.text.x = element_text(size=1, angle=90, hjust=0.5, vjust=0.5),
        legend.position = "none") +
  labs(title = "Dotplot between Date and Daily Total Counts",
       x = "Date",y = "Daily Total Counts") +
  facet_grid(rows = vars(StateAbbr), cols =  vars(Category))
```



We are looking for the daily attorney counts over time. According to the graph, we conclude that there are a generally positive linear relationship in the attorney daily total counts in all categories and in selected five states. The Family and Children related consultation is the most popular category in these states, whose daily total counts are obviously more than other categories overall. Specifically, the category of Family and Children consultation in all five states has the strongest positive linear relationship and the steepest slope among all categories, which means as one date passing, the daily total attorney counts in category of Family and Children will have the most increase of all categories in five states on average. Comparing them vertically, the linear relationship of FL is the strongest, then is the that of IL state, followed by state IN and SC with more flatter slopes, and the Family and Children daily total counts varied heavily over time in TX.The relationships between date and daily total counts are also strong in category of Housing and Homelessness, and category of Other, but are significantly weaker than that in the category of Family and Children. They share the similar patterns in five states. State FL has relatively stronger linear relationship between date and daily total counts of these two categories than other four states; State TX has insignificantly weaker linear relationship in these two categories; And for other states, the linear relationships are really weak. Besides, the daily total counts of “Other” category is slightly higher than that of “Housing and Homelessness” category on average. Last but not least, the linear relationships of remaining categories in five states are so weak that their slopes are almost horizontal, and the daily total counts of those remaining categories are low and no more than 20 on average. In particular, juvenile is the most unpopular categories; there are only few counts of consultation about juvenile in four states and no counts about it in state SC. By the way, the state TX only has few counts about Education category. It is pretty interesting that although the States of IN and SC provide the consultation for a longer time than other states, but the linear relationships between date and daily total counts of all categories are more stronger on average in FL and IL.

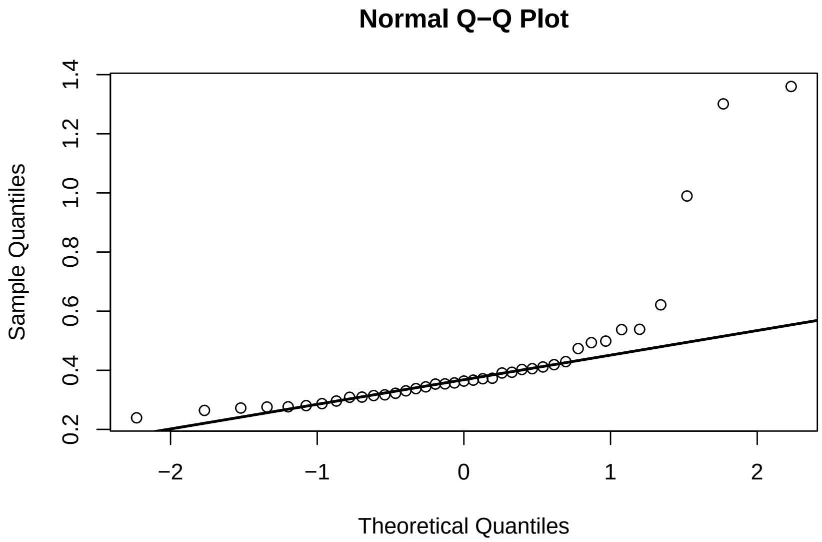
Research Question

We have illustrated the relationship between several variables. But we are still wondering what factors would affect the entries mean hours in every state, and are those independent variables significant to predict the attorney's quires time. Therefore, we want to find a best regression model for predicting times.

```
df_combined <- questions %>% mutate(ClosedOnUtc=as.Date(questions$ClosedOnUtc),
  AskedOnUtc=as.Date(questions$AskedOnUtc)) %>%
  group_by(StateAbbr) %>% mutate(total_questionCount = n()) %>% rowwise() %>%
  mutate(meanDuration = 24*as.numeric(mean(difftime(ClosedOnUtc, AskedOnUtc)), na.rm = TRUE)) %>%
  distinct(StateAbbr, .keep_all = TRUE) %>% select(StateAbbr, total_questionCount, meanDuration)
df_combined <- right_join(df_combined, attorneytimeentries, by = join_by(StateAbbr == StateAbbr))
df_combined <- df_combined %>% group_by(StateAbbr) %>% mutate(total_attorneyCount = n()) %>%
  mutate(meanHour = mean(Hours))%>% select(-Id, -TimeEntryUno, -AttorneyUno, -Hours) %>%
  distinct(StateAbbr, .keep_all=TRUE)
```

Here, we modified the original dataset and selected the dependent variables needed.

```
a<- qqnorm(df_combined$meanHour)
qqline(df_combined$meanHour,  lwd = 2)
```



We assume the dependent variable, mean hour, is normal as the majority of these points follows the straight dashed line but the QQ plots shows that there are a lot of outliers. So, we decide to do two groups of analysis either with or without outliers.

With Outliers

```
m1 <- lm(meanHour ~ total_attorneyCount + total_questionCount + meanDuration, data = df_combined)
summary(m1)
```

```
##
## Call:
## lm(formula = meanHour ~ total_attorneyCount + total_questionCount +
##     meanDuration, data = df_combined)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.23149 -0.10887 -0.07199  0.00278  0.88197
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.750e-01  5.692e-02   8.345 7.68e-10 ***
## total_attorneyCount -2.587e-05  2.574e-05  -1.005    0.322
## total_questionCount  8.454e-06  1.438e-05   0.588    0.560
## meanDuration        -4.852e-06  9.508e-06  -0.510    0.613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2528 on 35 degrees of freedom
## Multiple R-squared:  0.04112,    Adjusted R-squared:  -0.04108
## F-statistic: 0.5002 on 3 and 35 DF,  p-value: 0.6846
```

```r
m2 <- lm(meanHour ~ total_attorneyCount + total_questionCount, data = df_combined)
m3 <- lm(meanHour ~ total_attorneyCount, data = df_combined)
compareLM(m1, m2, m3) # model 3 prefer
```

```
## $Models
##   Formula
## 1 "meanHour ~ total_attorneyCount + total_questionCount + meanDuration"
## 2 "meanHour ~ total_attorneyCount + total_questionCount"
## 3 "meanHour ~ total_attorneyCount"
##
## $Fit.criteria
##   Rank Df.res   AIC   AICc   BIC R.squared  Adj.R.sq p.value Shapiro.W
## 1    4     35 9.192 11.010 17.51   0.04112 -0.041080  0.6846    0.6511
## 2    3     36 7.481  8.658 14.14   0.03398 -0.019690  0.5367    0.6379
## 3    2     37 5.711  6.396 10.70   0.02829  0.002023  0.3061    0.6480
##   Shapiro.p
## 1 2.217e-08
## 2 1.421e-08
## 3 1.996e-08
```

We constructed three models to predict the mean attorney times. We assumed the mean attorney times in each state is influenced by the variables, total attorney count in a state, total question counts asked by clients in a state, and the mean attorney duration. So we built the first model of predicting the mean hour by all these three independent variables. But from the results of first model, the Pr(>|t|) values in the t-test for independent variables are highly more than the 5% or 10% significance level, which means they are not significant to predict the mean hour. So we excluded the most insignificant variable, mean duration, whose Pr(>|t|) value is 0.613, which the highest value in these three variables. Similarly, the independent variable of total question count, whose Pr(>|t|) value is second highest and equal to 0.560, is excluded from the third model as well.

Then, we compared three models to find out the best model. The model 3, which represents the mean hour is only predicted by the total attorney count, has the lowest AIC(5.711 compared to 7.481 of model 2 and 9.192 of model1) and BIC(10.7 compared to 14.14 of model 2 and 17.51 of model1) and the highest adjust R-square, which is 0.002023 compared to -0.041080 and -0.019690); so we prefer model 3 is the best model to predict the mean hour. But we have already known there is a positive relationship between the hour and attorney count in our visualization, additionally, these three independent variables are not significant predictors. Based on the conclusion and the QQ-plot above, we believed there must be some outliers influence the model to predict the mean attorney hours, so we built three same models but without outliers to make better prediction.

Without Outliers

```
Lower <- quantile(df_combined$meanHour)[1] - 1.5*IQR(df_combined$meanHour)
Upper <- quantile(df_combined$meanHour)[2] + 1.5*IQR(df_combined$meanHour)
df_nooutlier <- subset(df_combined, df_combined$meanHour > Lower & df_combined$meanHour < Upper)
```

Here, we removed the outliers from the dataset.

```
m1 <- lm(meanHour ~ total_attorneyCount + total_questionCount + meanDuration, data = df_nooutlier)
summary(m1)
```

```
##
## Call:
## lm(formula = meanHour ~ total_attorneyCount + total_questionCount +
##     meanDuration, data = df_nooutlier)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.110231 -0.032126  0.002683  0.028744  0.122491
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.452e-01  1.392e-02  24.809   <2e-16 ***
## total_attorneyCount -1.294e-05  5.686e-06  -2.275   0.0311 *
## total_questionCount  7.285e-06  3.364e-06   2.165   0.0394 *
## meanDuration        -6.990e-07  2.103e-06  -0.332   0.7422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05402 on 27 degrees of freedom
## Multiple R-squared:  0.1694, Adjusted R-squared:  0.07716
## F-statistic: 1.836 on 3 and 27 DF,  p-value: 0.1644
```

```
m2 <- lm(meanHour ~ total_attorneyCount + total_questionCount, data = df_nooutlier)
m3 <- lm(meanHour ~ total_attorneyCount, data = df_nooutlier)
compareLM(m1, m2, m3) # model 2 prefer
```

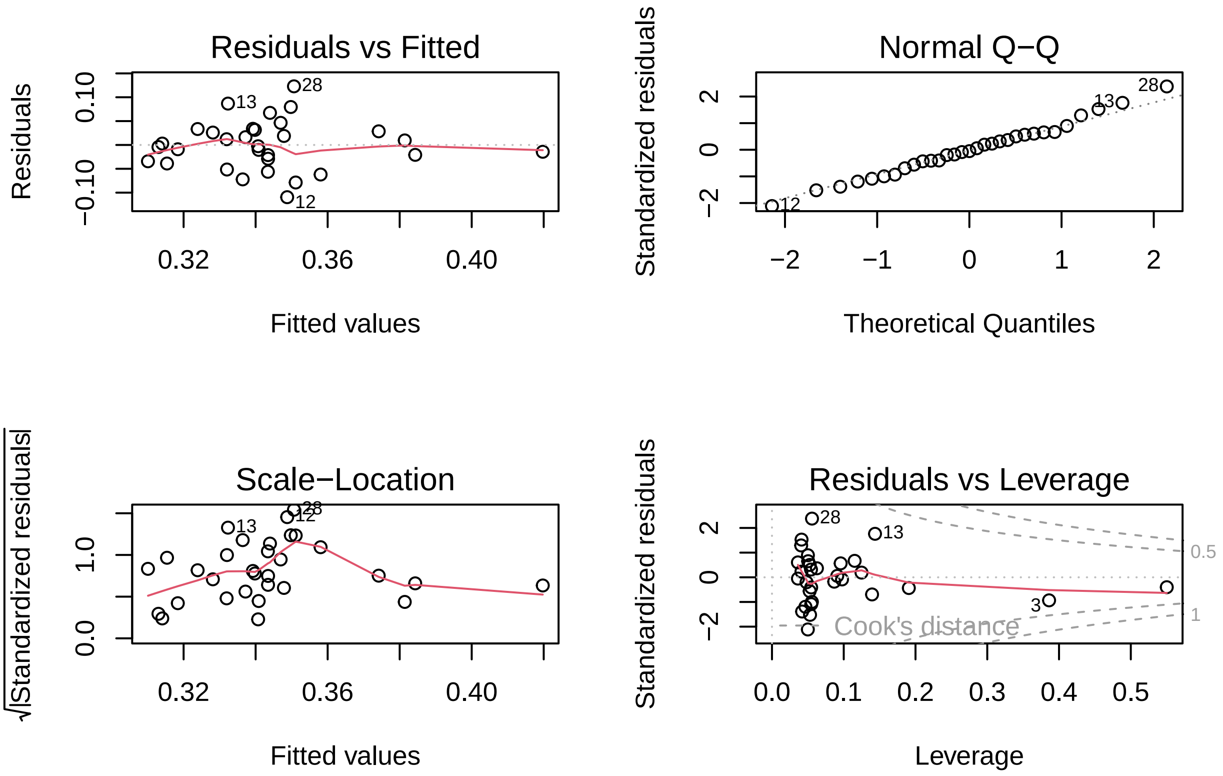
```
## $Models
##   Formula
## 1 "meanHour ~ total_attorneyCount + total_questionCount + meanDuration"
## 2 "meanHour ~ total_attorneyCount + total_questionCount"
## 3 "meanHour ~ total_attorneyCount"
##
## $Fit.criteria
##   Rank Df.res    AIC   AICc    BIC R.squared Adj.R.sq p.value Shapiro.W
## 1    4     27 -87.25 -84.85 -80.08   0.16940  0.07716 0.16440    0.9918
## 2    3     28 -89.12 -87.58 -83.39   0.16600  0.10650 0.07871    0.9921
## 3    2     29 -86.14 -85.25 -81.84   0.02067 -0.01310 0.44040    0.9909
##   Shapiro.p
## 1    0.9968
## 2    0.9976
## 3    0.9943
```

```
summary(m2)
```

```
##
```

```
## Call:
## lm(formula = meanHour ~ total_attorneyCount + total_questionCount,
##     data = df_nooutlier)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109533 -0.031610 -0.002707  0.030050  0.122860
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.446e-01  1.358e-02  25.385   <2e-16 ***
## total_attorneyCount -1.250e-05  5.445e-06  -2.296   0.0294 *
## total_questionCount  6.926e-06  3.135e-06   2.209   0.0355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05316 on 28 degrees of freedom
## Multiple R-squared:  0.166,  Adjusted R-squared:  0.1065
## F-statistic: 2.787 on 2 and 28 DF,  p-value: 0.07871
```

```
par(mfrow = c(2, 2))
plot(m2)
```



Similarly, we summarized the model 1 firstly, and found the Pr(>|t|) values of two variables, total attorney count and total question count, are 0.0311 and 0.0394 respectively and smaller than the default 5% significance level, which indicate they are significant predictors. But duration variable has 0.7422 Pr(>|t|) value, we believed it is not a really great predictor. And after contrasting three models, we preferred model 2 to predict the mean hour as it had the highest adjust R-square, which is 0.1065, with approximately AICs and BICs.

So we finally determined the model 2 without outliers, including total attorney counts and total question counts as independent variables, is the best model to predict the mean attorney hours in a state. And we did another summary for model 2.

The least-square estimate for the regression model of model 2 is estimate mean hour = 3.446e-01 - 1.250e-05 * (total attorney counts) + 6.926e-06 * (total question count). That means a unit increase in total attorney counts will cause a unit decrease in mean attorney hour by 1.250e-05 hours, holding all other variables in regression constant; a unit increase in total question counts will cause a unit increase in mean attorney hour by 6.926e-06, holding all other variables in regression constant. When the total attorney count and total question count are equal to zero, we expect the mean attorney hour is 3.446e-01.

The residual standard error is 0.05316 on 28 degrees of freedom. Moreover, the adjust R-square is 0.1065, which means the that approximately 10.65% of the variability in mean attorney hours is explained by the independent variables,total attorney and question counts, in the regression model. By the way, the p-value in F test is 0.07871, which also prove two independent variables quite significant to predict the outcome variable mean attorney hours at 10% significance level; more importantly, Pr(>|t|) values of variables total attorney counts and total question counts are 0.0294 and 0.0355, smaller than 5% level of significance , which indicate these two independent variables are signifcant to predict the mean hours.

Residual analysis for model 2:

Normality: Looking at the QQ plot, the majority of points follows the straight dashed line, so the assumption of normality is fulfilled.

Multicollinearity: the function vif() help to check. Multicollinearity for each variable is below 5, so the assumption of Multicollinearity is fulfilled.

Autocorrelation: Since the p-value = 0.3696 in the Durbin-Wastson test, we fail to reject the null hypothesis. Therefore we can assume that residuals are not auto-correlated.

Linearity: Looking at the Residuals vs Fitted graph, the residuals are spread randomly around the horizontal line and there is no particular pattern. Therefore, we can assume that linearity assumption is not violated.

Homoscedasticity: Looking at the Scale-Location plot, the residuals are spread randomly, and I can see a generally horizontal line with randomly spread points, so the assumption of homoscedasticity is fulfilled.

Outlier detection: there are no extreme values outside of the dashed line in the Residuals vs Leverage graph, so there are no influential values that might influence the regression results.

Summary / Conclusion

In this project, we began by exploring the dataset and visualizing some of the relationships between variables. The first visualization we created was a bar graph of total count of questions in each category, which helped us to understand the ranking of the categories. We then wanted to see how this ranking would vary by states, and we chose to focus on the top 5 states that have the most question inquiries. From the bar graph, we could see there are definitely some correlations between states and category, but not a fixed pattern. Next, we looked for whether there is correlation between total hours of conversations in each state, and the entries count in each state. And for the last visualization, we investigated in the daily attorney counts over time

After looking at these visualizations, we moved on to building models that would help us predict entries mean hours in every states. We found that the variable is approximately normal but has many outliers. So, we decided to do two analysis of regression models, either with or without outliers, to best determine the model that fits the data.

We constructed three models to predict the mean attorney times, with the variables, total attorney count in a state, total question counts asked by clients in a state, and the mean attorney duration. The first model consisted all three independent variables, all independent variables show a p-value that is insignificant at both the 5% and 10% significance level. So we excluded the most insignificant variable for the second and third model. We compared three models to find that model 3, where the mean hour is only predicted by the total attorney count, is the best model to predict the mean hour when including the outliers. But we believed there must be some outliers influencing the model and causing it to be less predictive, so we built the same three models but without outliers to make better prediction.

We determined the model 2 without outliers, including total attorney counts and total question counts as independent variables, is the best model to predict the mean attorney hours in a state. The least-square estimate for the regression model of model 2 is estimate mean hour = 3.446e-01 - 1.250e-05 * (total attorney counts) + 6.926e-06 * (total question count). It means a unit increase in total attorney counts will cause the mean attorney hour decrease by 1.250e-05 hours, and a unit increase in total question counts will cause increase in mean attorney hour by 6.926e-06, holding all other variables in regression constant. When the total attorney count and total question count are equal to zero, we expect the mean attorney hour is 3.446e-01. Approximately 10.65% of the variability in mean attorney hours is explained by the independent variables. Lastly, we checked that the assumptions for multicollinearity, normality, autocorrelation, and linearity to be met.

We had many key takeaways about this data after creating visualizations. When looking at the count of question asked in each state and the categories these questions are in, there are definitely some correlations between states and category, but not a fixed pattern. There is also a strong correlation between total hours of conversations in each state and the entries count in each state, which means length of conversations is related to the number of meetings made. It is also an unique project experience because during the process of building regression models, we found that outliers can have great influence in predicting the dependent variable. It is an important decision to make in whether to remove the outliers or not, in order to get the most accurate result. We can perform multiple analysis and compare each result to help us make the best decision. In the end, we saw how removing the outliers significantly reduced the p-values of some of the independent variables, making those variables significant predictors of the dependent variable.

Some major limitations that we faced was that the data were separted in several different datasets. We spot many potential relationships between variables, but most were recorded in separate datasets. We were able to rely on some tools and techniques to combine them into the same dataset. However, there were also problems that were less solvable. When looking for correlations between variables, we wondered how long the questions in each category took. There was no clear record of that, so we weren't able to perform the best investigation into the data. The suggestion we have is to include more relevant information, such as the time for questions in each category, which helps us understand what categories seem to be more concerning for people, and what categories are less significant; and also exclude some of the irrelevant data, like the actual question board, which had an enormous amount of observations that gave not much meanings.