Predicting Total Damages Awarded to Plaintiffs in Civil Cases

Manish Nagireddy
mnagired

Due Weds, October 16, at 8:00PM

Contents

Introduction

Although everyone loves reading about Supreme Court cases, the common citizen's experience with the judicial system is usually with civil courts, if at all. Thus, when a qualm arises about the nature of damages awarded to the plaintiff, it must be taken with serious consideration as such decisions will likely impact most people's lives in some capacity at a point in time. This project, therefore, seeks to discover whether there exists a significant relationship between the total damages awarded to a plaintiff and a variety of factors, such as total amount of damages requested by the plaintiff, the length of trial, and the type of claim made by the plaintiff.

Exploratory Data Analysis

Data

In the court data, we analyze a random sample of 126 cases and 4 variables. Because of our interest in total damages awarded to the plaintiff, we look at the relationship between total damages and the three explanatory variables: total amount of damages demanded by the plaintiff, length of the trial, and type of claim made by the plaintiff.

`TOTDAM`: total amount of damages awarded to plaintiff (in $)

`DEMANDED`: total amount of damages requested from the court by plaintiff (in $)

`TRIDAYS`: how many days the trial lasted

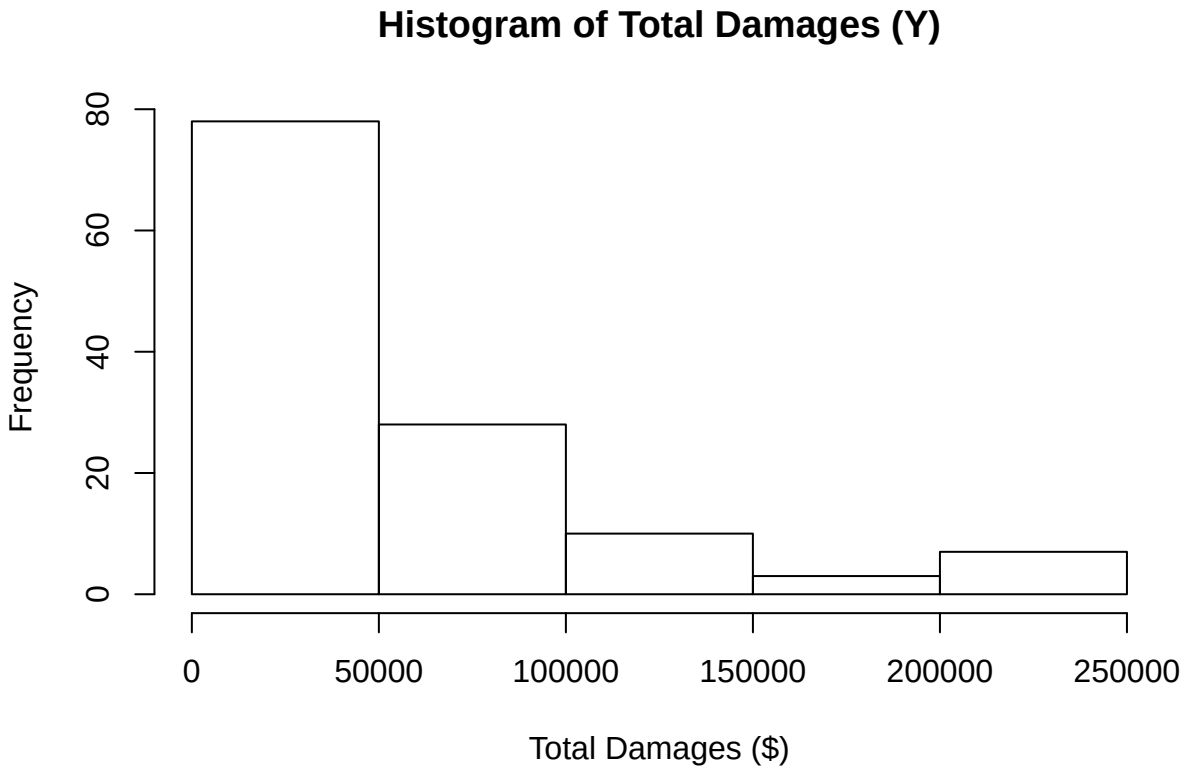
`CLAIMTYPE`: type of claim the plaintiff made - *categorized as follows:* 1: motor vehicle; 2: premises liability; 3: malpractice; 4: fraud; 5: rental/lease ; 6: other

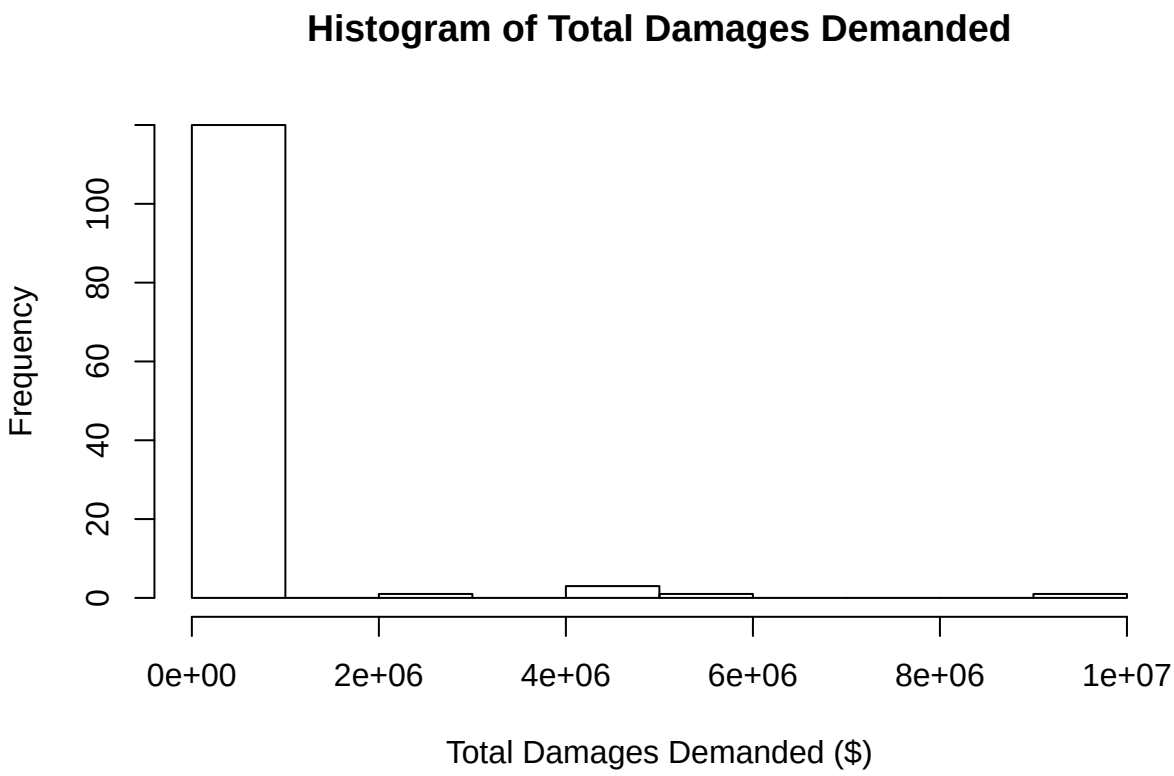
The first few lines of data appear as follows:

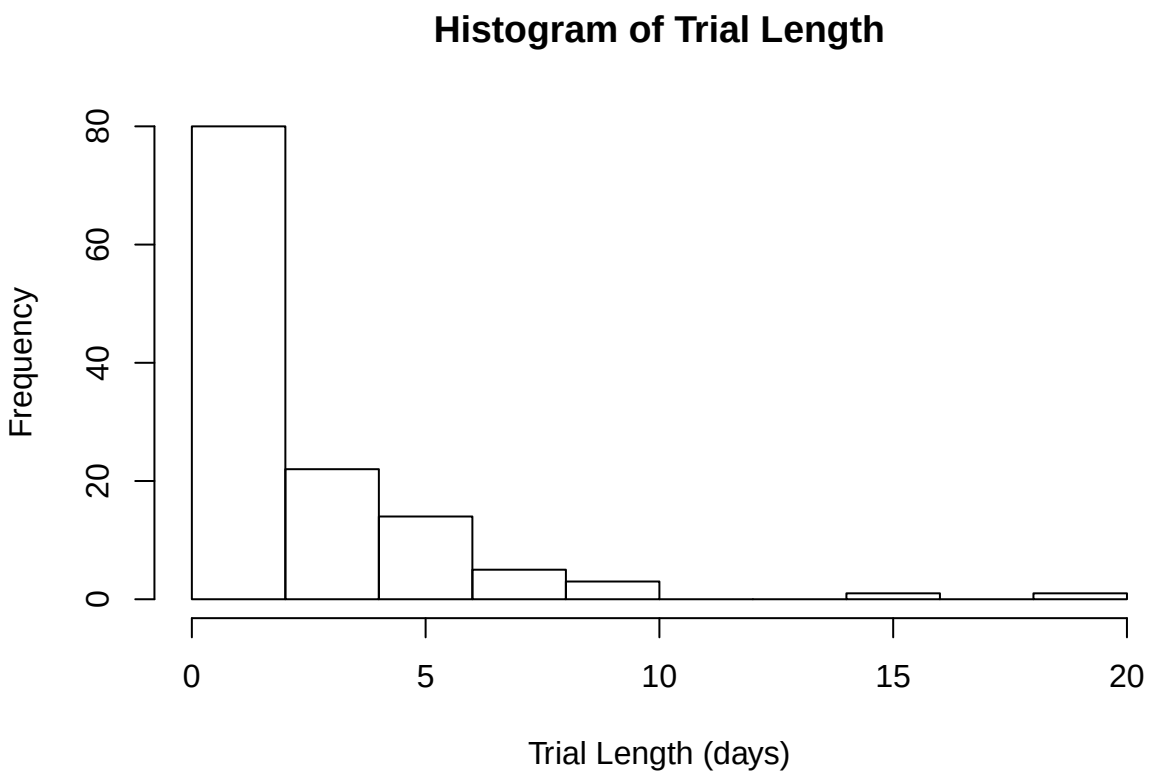
```
## # A tibble: 10 x 4
##    TOTDAM DEMANDED TRIDAYS CLAIMTYPE
##     <int>    <int>   <int> <fct>
##  1  11760    17640       1 Rental
##  2 150000   200000       2 Other
##  3   2831     2870       1 Other
##  4  29863     9900       5 Motor
##  5   2200     2200       2 Other
##  6  70945    58816       2 Other
##  7   2319    10000       1 Other
##  8 141512   121110       1 Other
##  9  82780   780382       5 Other
## 10   7245    21196       3 Other
```

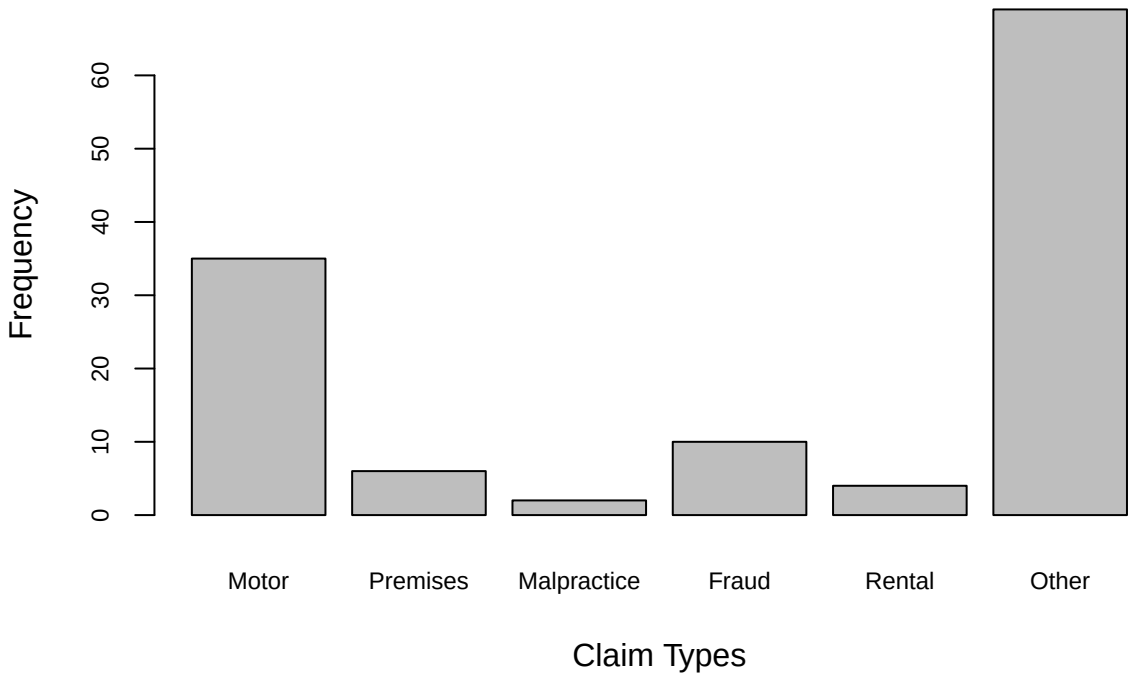
Univariate Exploration

A common first step in analysis is to explore each variable individually. We use histograms to explore the distribution of our quantitative (continuous) variables (total damages, total damages demanded, trial days) and a barplot to describe our categorical variable (claim type).


Histogram of Total Damages (Y)
Frequency
0
20
40
60
80
0
50000
100000
150000
200000
250000
Total Damages ($)


Histogram of Total Damages Demanded
Frequency
0
20
40
60
80
100
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Total Damages Demanded ($)


Histogram of Trial Length
Frequency
0
20
40
60
80
0
5
10
15
20
Trial Length (days)



We supplement the univariate graphical summaries with the following numerical summaries:

For Total Damages:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     225    7544   26322   51279   70750  248280
```

```
## [1] 59746.78
```

For Total Damages Demanded:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1435    21272    50000   357088   132500 10000000
```

```
## [1] 1249670
```

For Length of Trial:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.833   4.000  20.000
```

```
## [1] 2.830548
```

For Claim Type:

```
##       Motor    Premises Malpractice       Fraud      Rental       Other
##          35           6           2          10           4          69
```

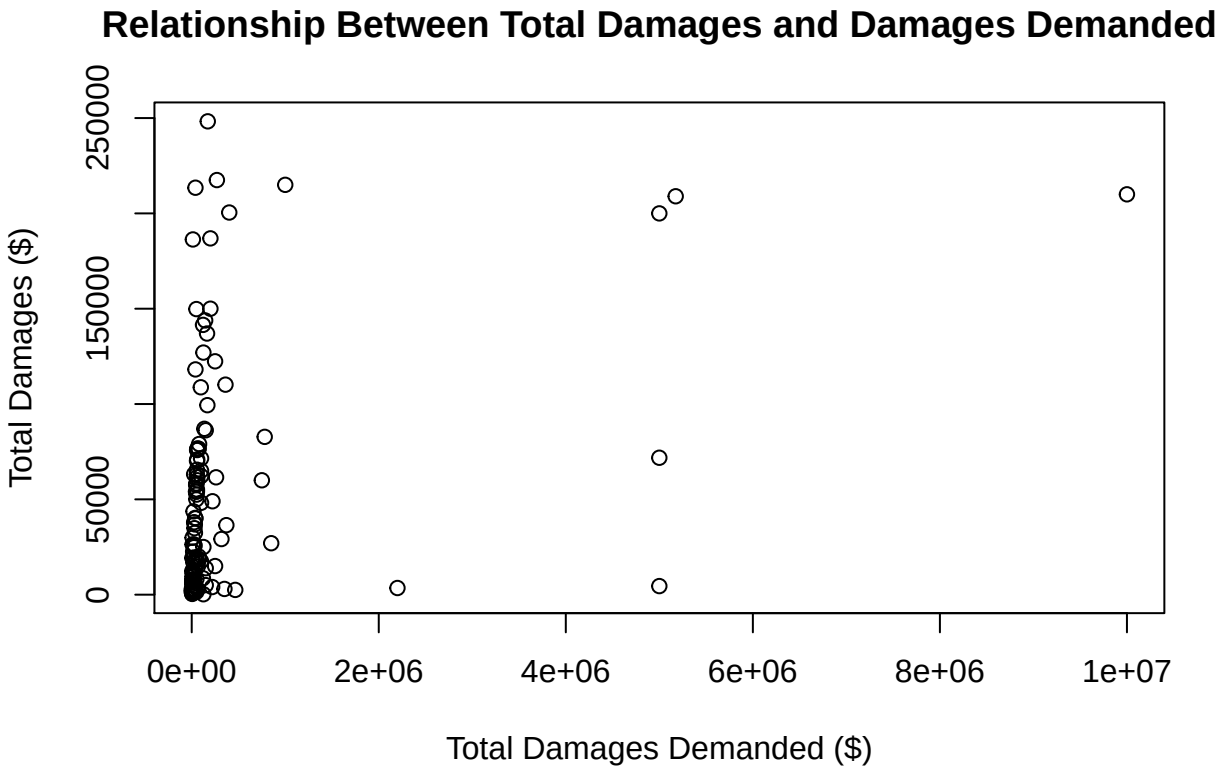
After looking at both the graphs and the summary statistics of our four variables, we can make the following observations:

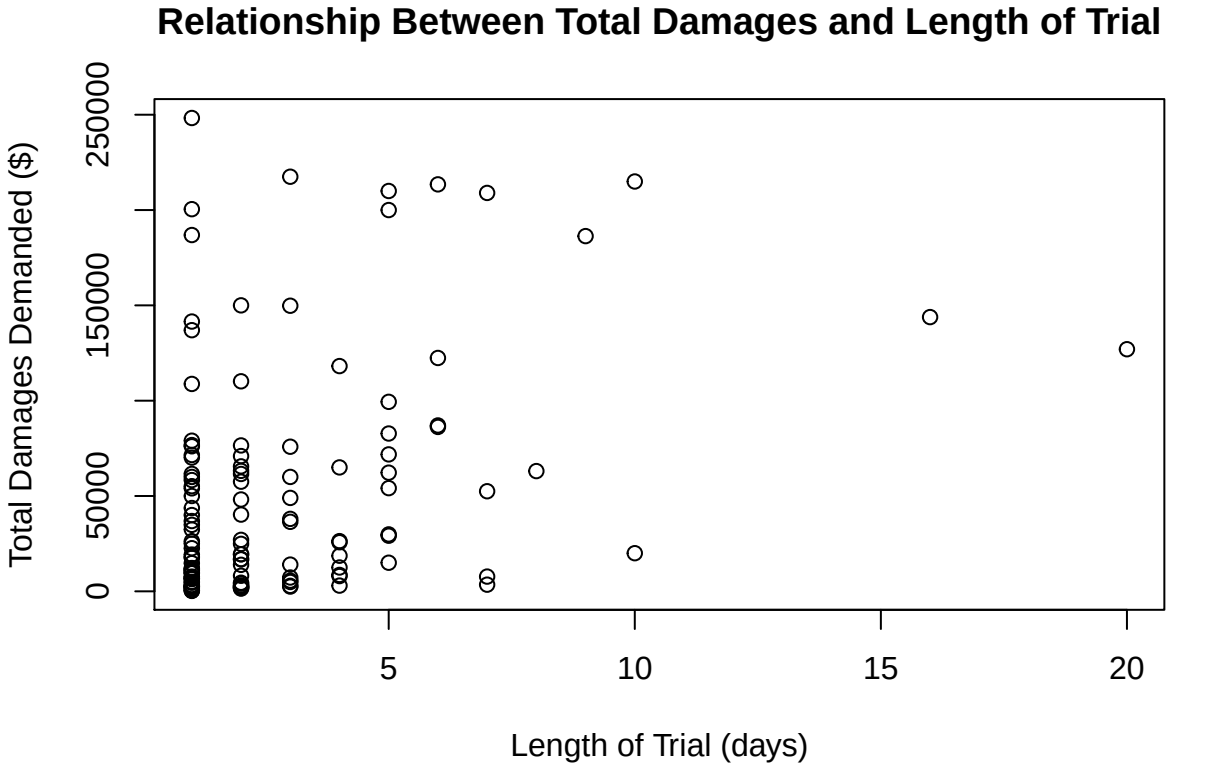
The distribution of total damages (`TOTDAM`) is clearly right skewed and unimodal, with a mean of $51,279. The distribution of total damages demanded (`DEMANDED`) is also clearly right skewed, but the graph becomes

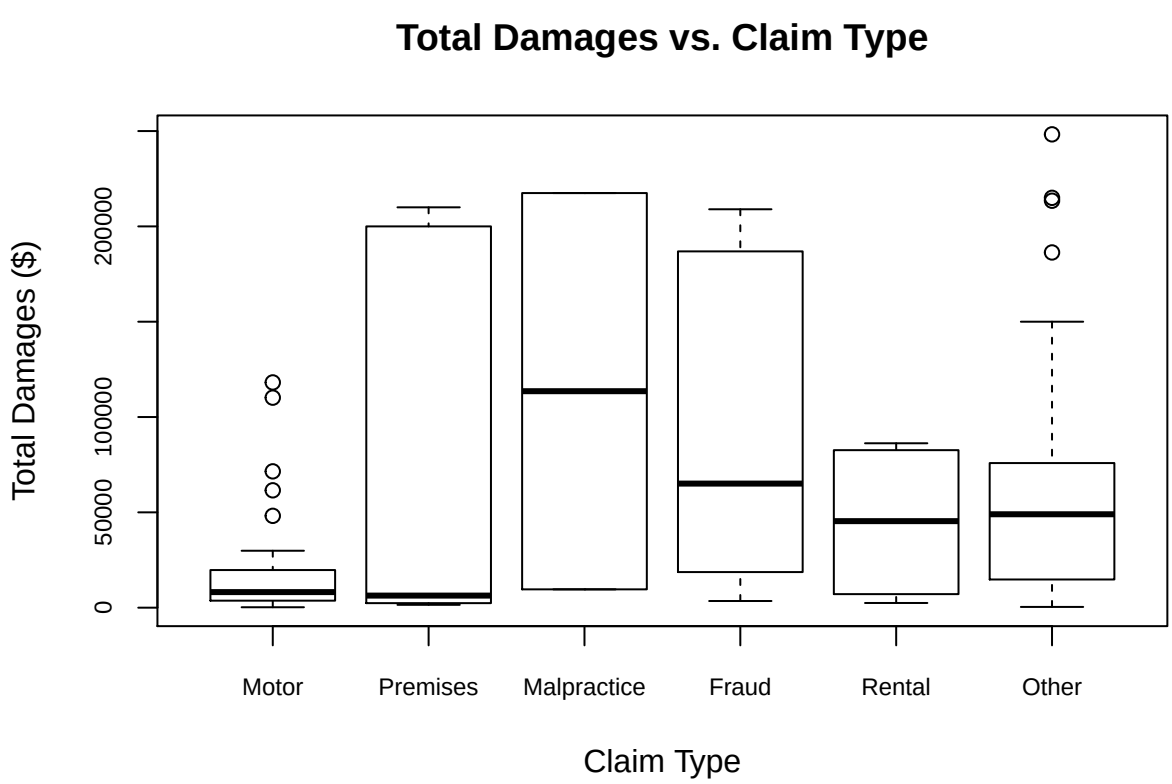
distorted because of the data point in row 99, with a value of 10000000 for the total damges demanded. The distribution of trial length (`TRIDAYS`) is also right skewed, as most trials are under 10 days. Looking at our only categorical variable (`CLAIMTYPE`), we see the types of claims distributed in the following manner: 35 motor, 6 premises, 2 malpractice, 10 fraud, 4 rental, and 69 other types of claims.

Bivariate Exploration

Now that we understand the distributions of the variables in our data, we can look graphically at how each predictor is associated with the response variable, total damages (`TOTDAM`).

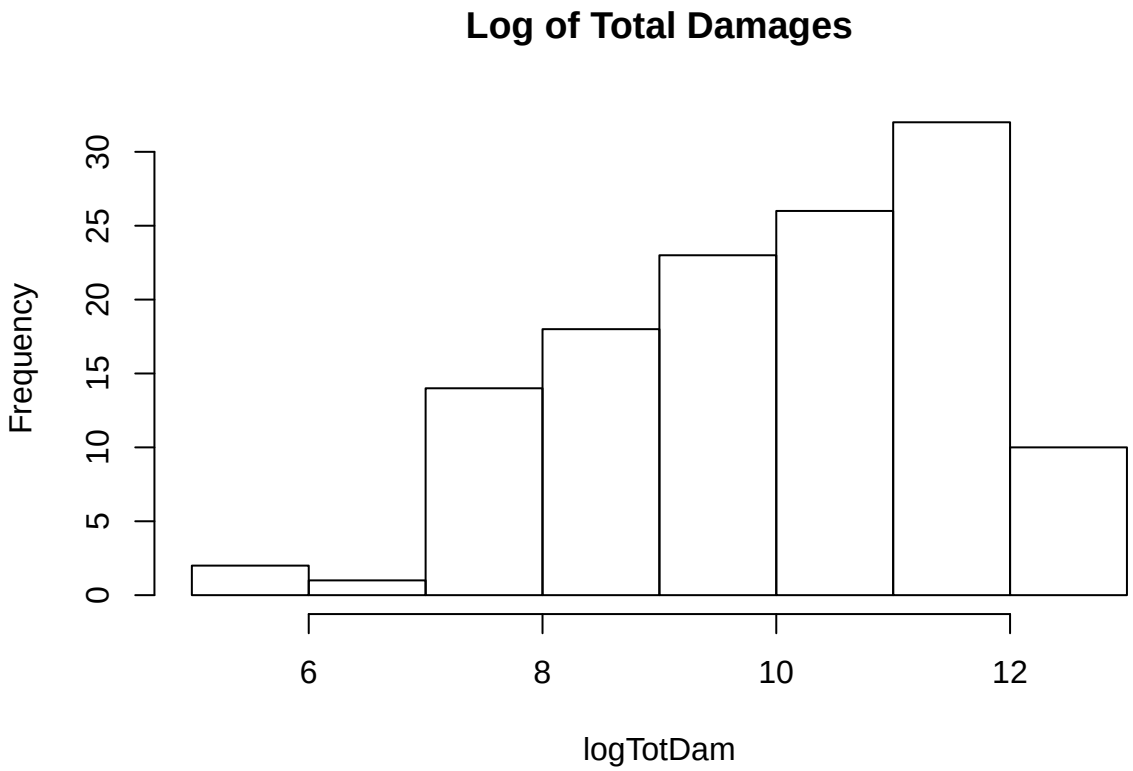

Relationship Between Total Damages and Damages Demanded
Total Damages ($)
0
50000
150000
250000
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Total Damages Demanded ($)


Relationship Between Total Damages and Length of Trial
Total Damages Demanded ($)
0
50000
150000
250000
5
10
15
20
Length of Trial (days)

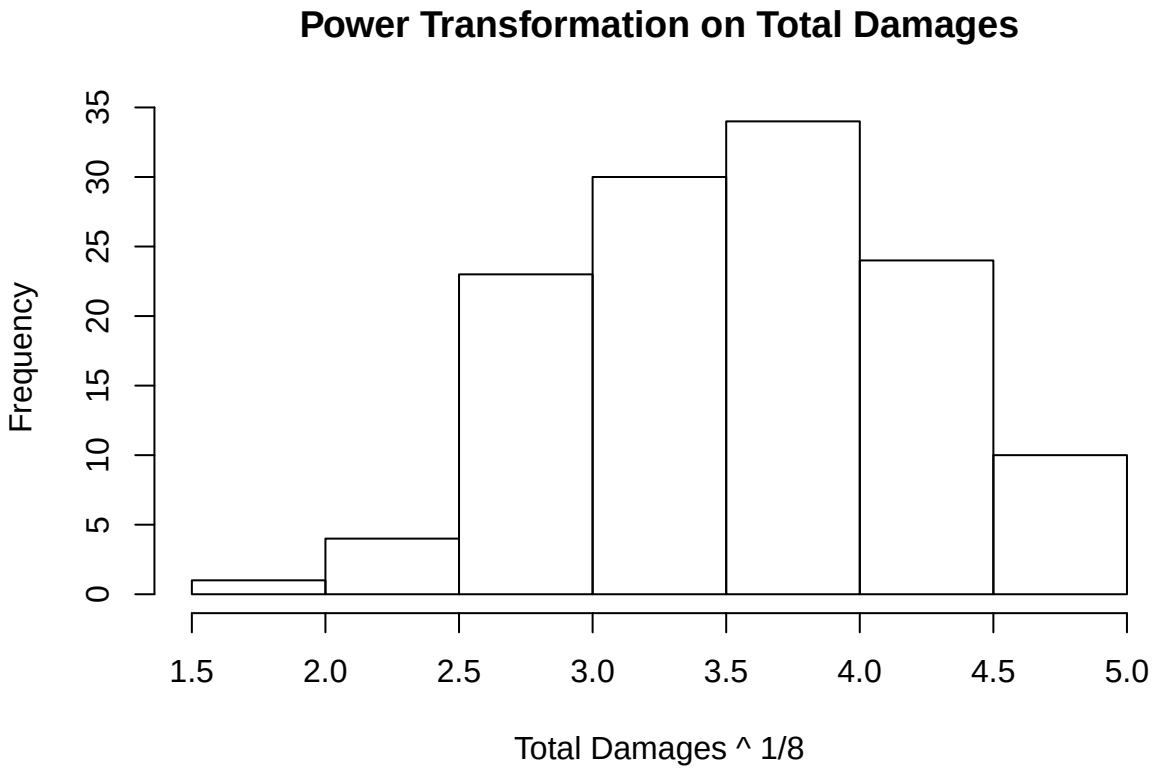


When exploring the bivariate data, we see that the total damages does not appear to have a strong linear association with either of the quantitative variables. Perhaps, at best, total damages has an extremely weak negative relationship with the length of the trial. This can likely be explained by the fact that the distribution of total damages is staggeringly skewed right, meaning we would have to try some power or log transformations. Additionally, there does not seem to be enough of a difference in total damages between the types of claims to justify hypthesizing that claim type significantly affects total damages (but again we need to transform the response variable before making any sound judgments).

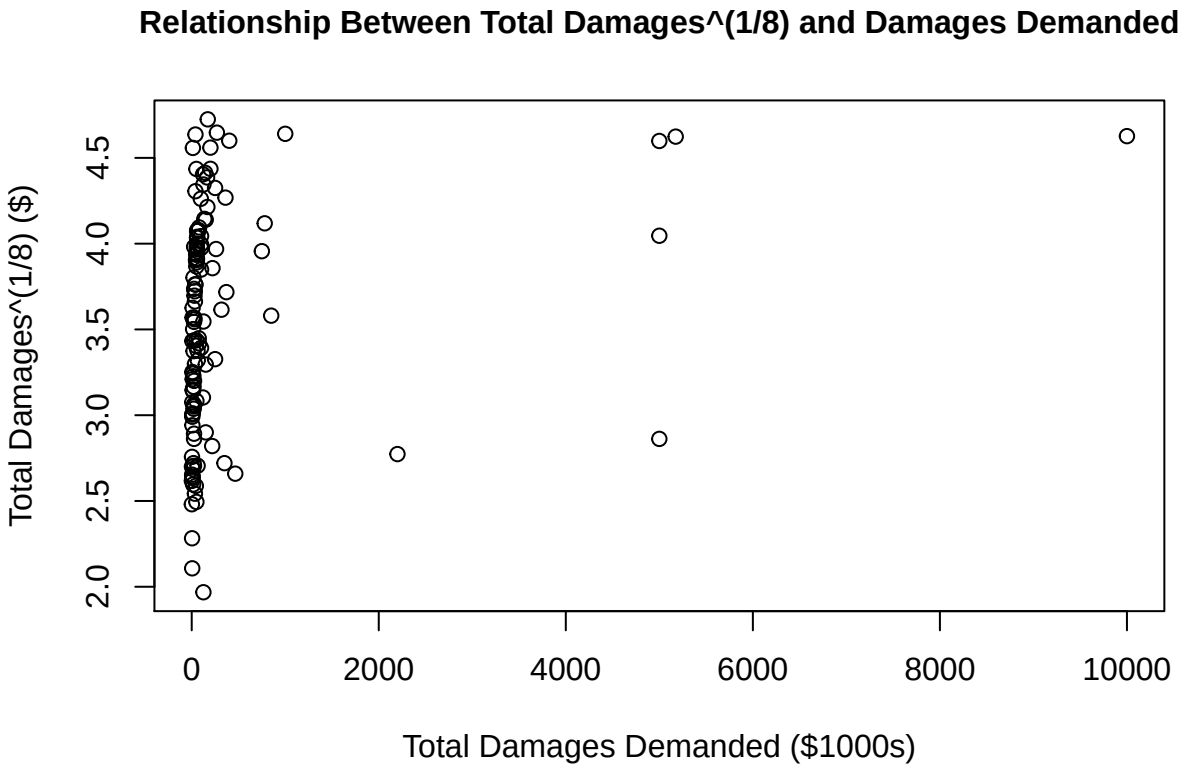
Modeling

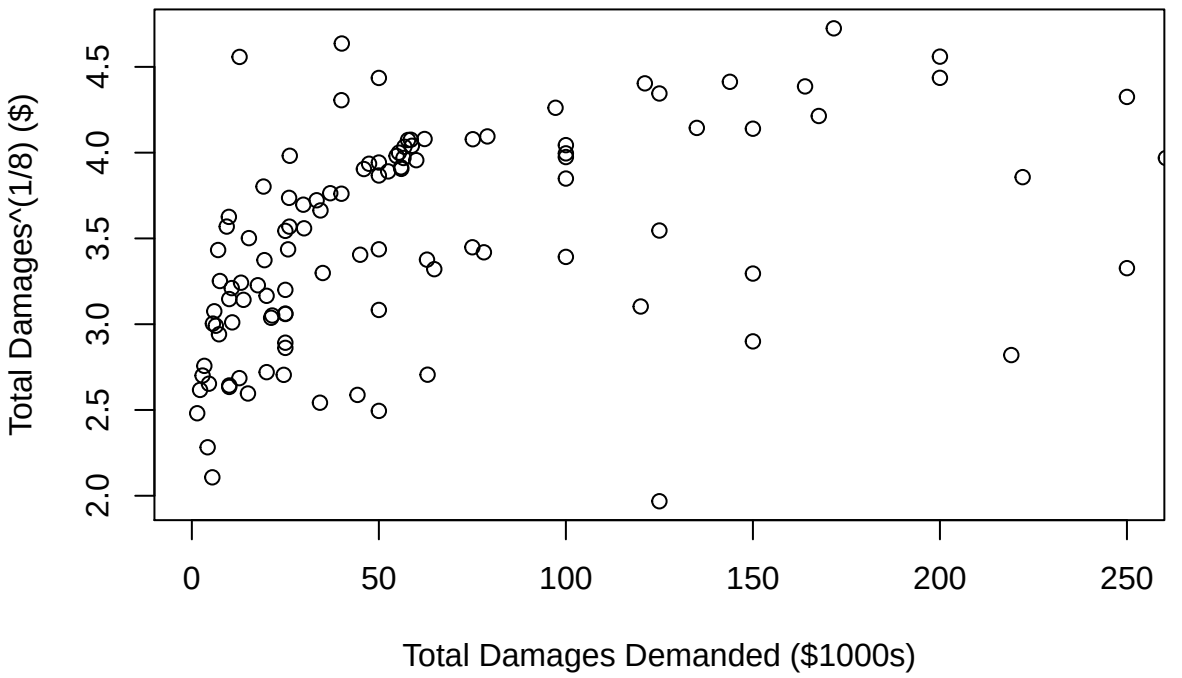


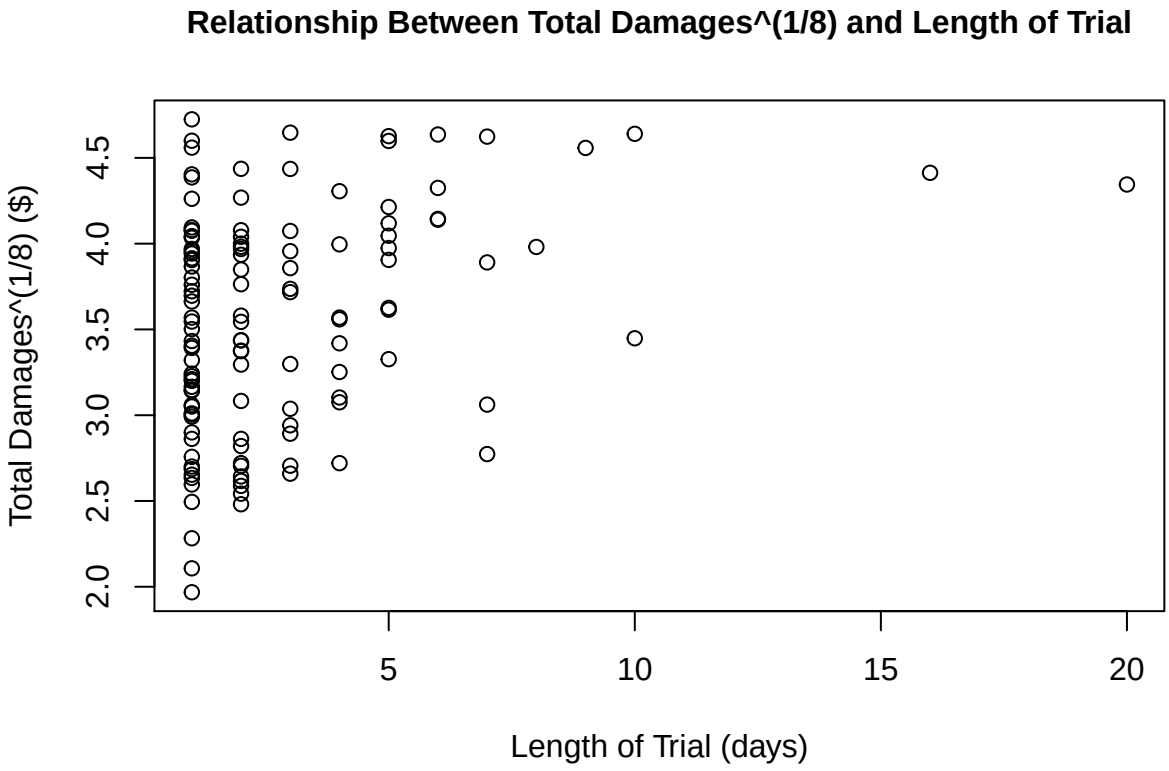
If we take the log of total damages, the resulting distrbution becomes skewed left which means we cannot use it.

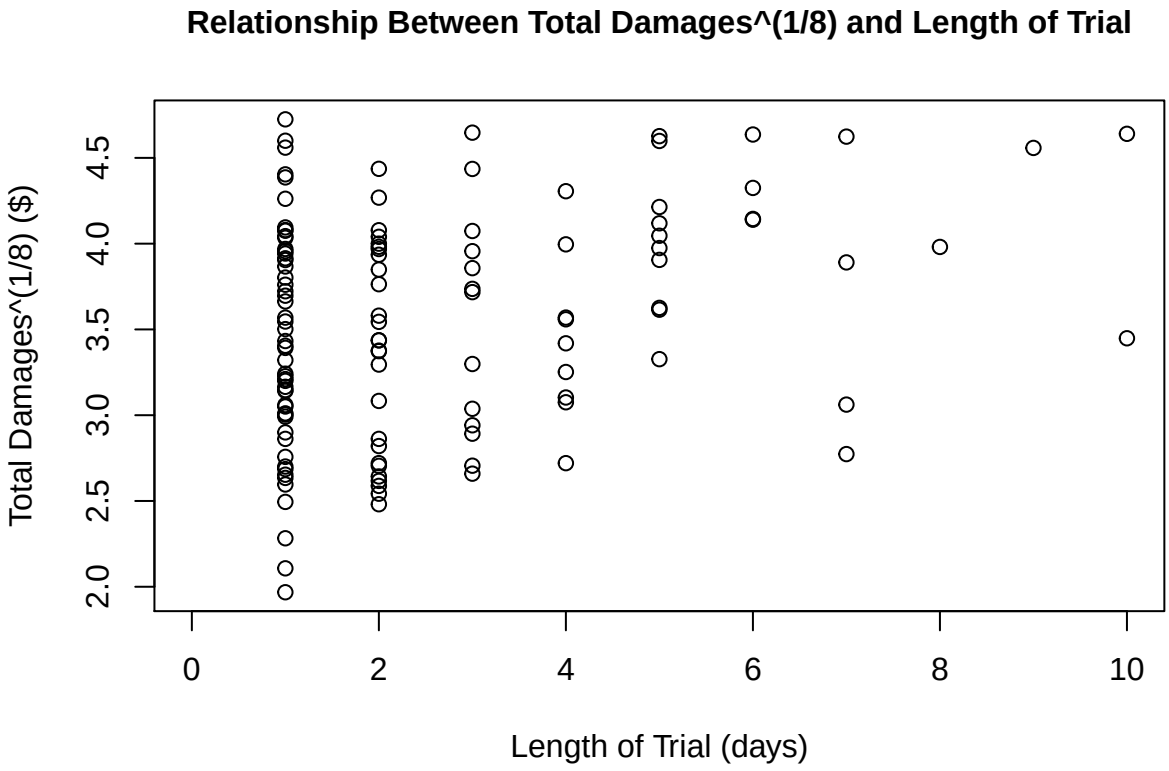


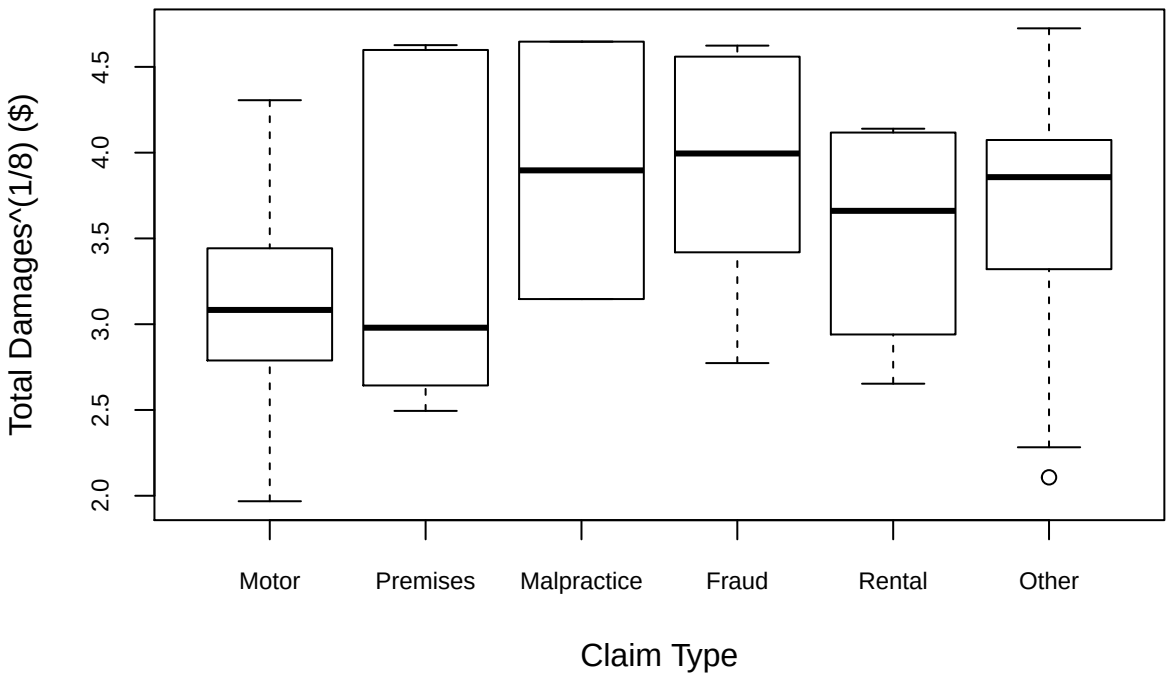
If we raise the response variable to the 8th, it appears that we have a reasonably symmetric distribution. Although this will make the interpretations of our results rather hard to apply, the ability to use reasonably symmetric data is well worth the tradeoff.


Relationship Between Total Damages^(1/8) and Damages Demanded
Total Damages^(1/8) ($)
2.0
2.5
3.0
3.5
4.0
4.5
0
2000
4000
6000
8000
10000
Total Damages Demanded ($1000s)


Relationship Between Total Damages^(1/8) and Damages Demanded
Total Damages^(1/8) ($)
2.0
2.5
3.0
3.5
4.0
4.5
0
50
100
150
200
250
Total Damages Demanded ($1000s)


Relationship Between Total Damages^(1/8) and Length of Trial
Total Damages^(1/8) ($)
2.0
2.5
3.0
3.5
4.0
4.5
5
10
15
20
Length of Trial (days)

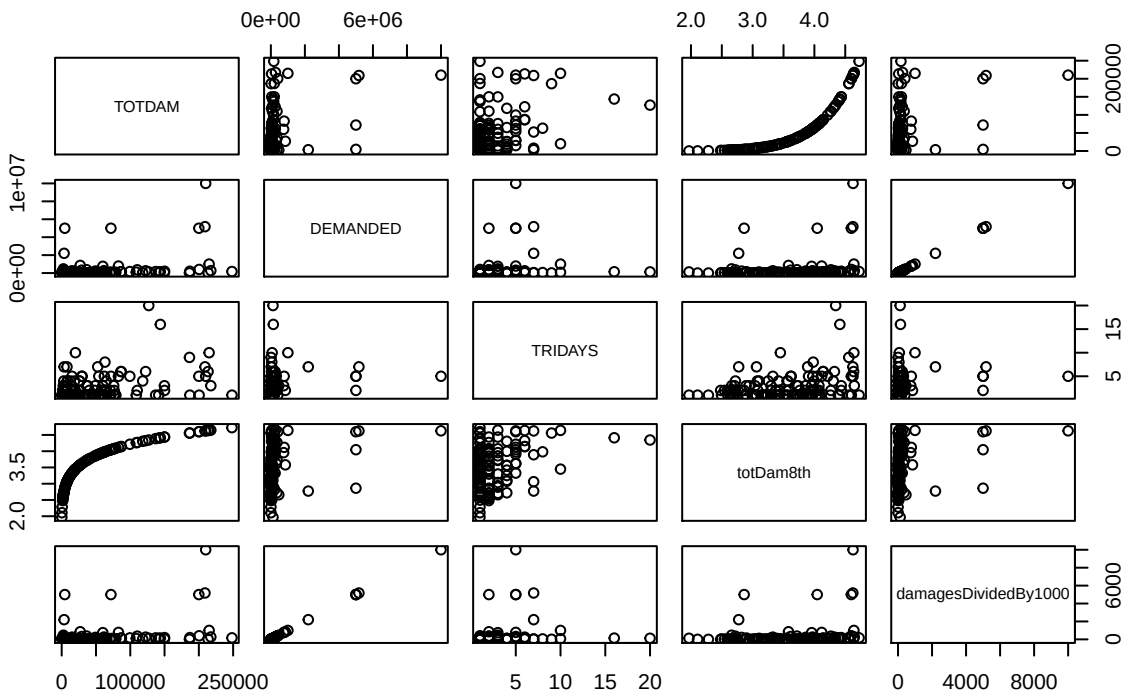

Relationship Between Total Damages^(1/8) and Length of Trial
Total Damages^(1/8) ($)
2.0
2.5
3.0
3.5
4.0
4.5
0
2
4
6
8
10
Length of Trial (days)



After raising the response variable to the 1/8th power, we can see some associations with the response variables emerging. When we limit the range of the Damages Demanded to $250 (thousand), we see a weak positive linear relationship with (total damages) ˆ 1/8. By limiting the length of the trial to 10 days, we can also see a weak positive relationship with the (total damages) ˆ 1/8. Although it doesn't jump out to us immediately, we cannot disregard the categorical variable (claim type) as it may have an association with the (total damages)ˆ1/8th power.

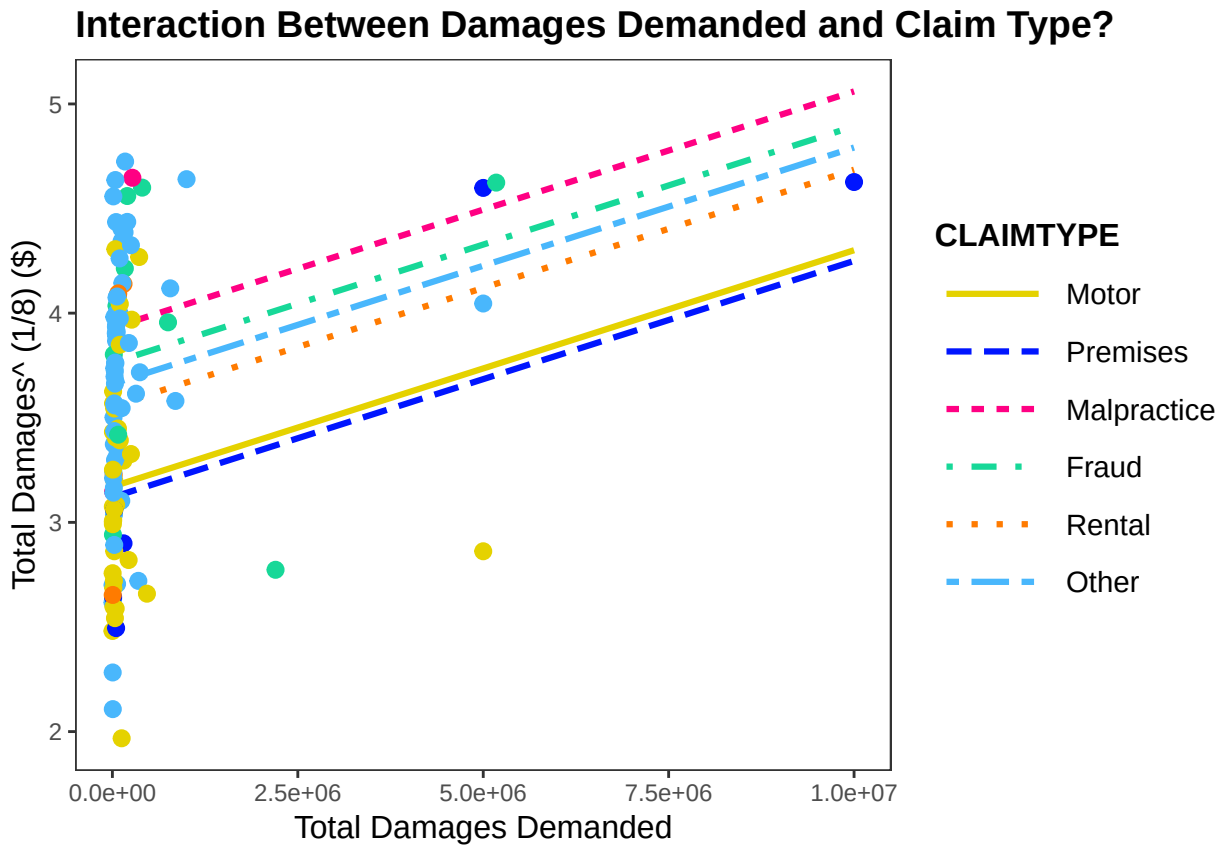
Therefore, all variables may be useful in this model, but we need to first check for multicollinearity. An indication of possible multicollinearity is when we get relatively strong correlations between pairs of explanatory variables and for this we use the pairs plot as well as a correlation matrix.

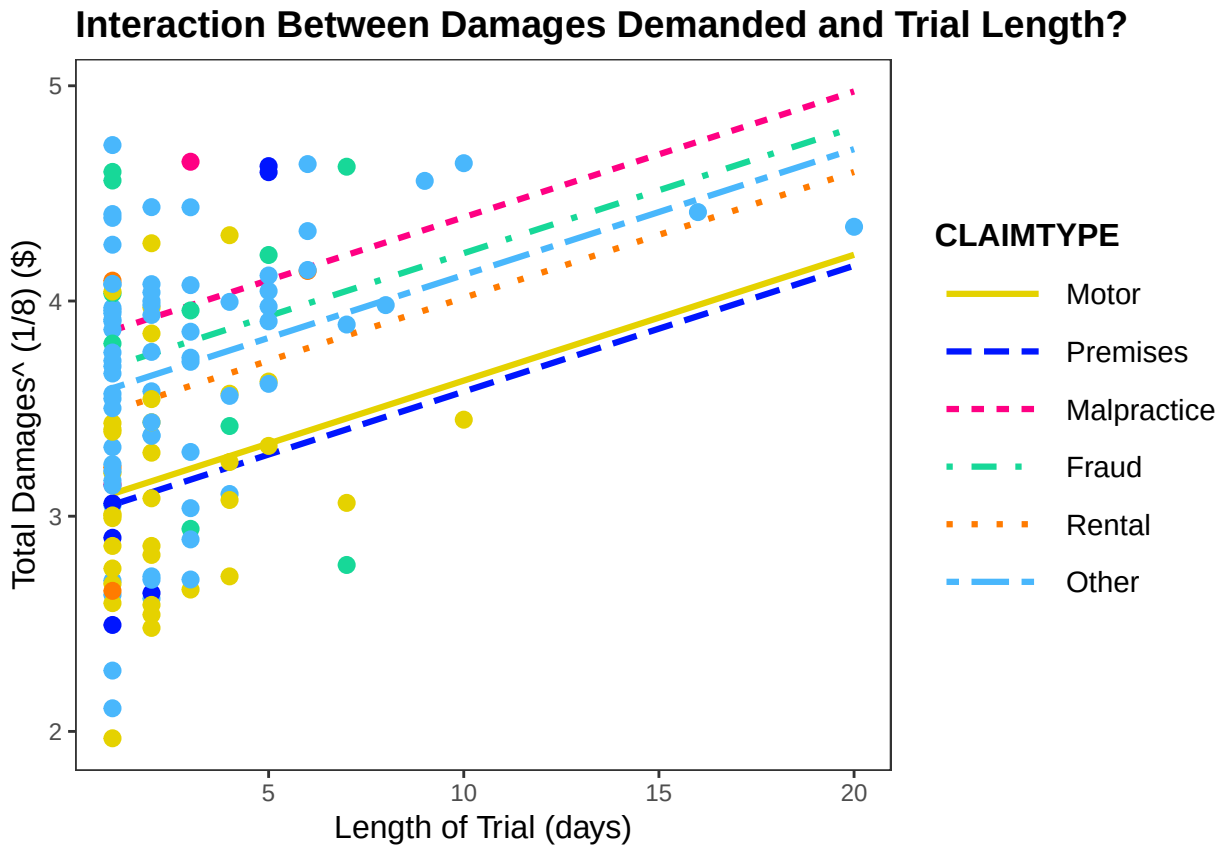
```
##                         TOTDAM DEMANDED TRIDAYS totDam8th
## TOTDAM                    1.00     0.35    0.35      0.86
## DEMANDED                  0.35     1.00    0.18      0.22
## TRIDAYS                   0.35     0.18    1.00      0.33
## totDam8th                 0.86     0.22    0.33      1.00
## damagesDividedBy1000      0.35     1.00    0.18      0.22
##                         damagesDividedBy1000
## TOTDAM                                  0.35
## DEMANDED                                1.00
## TRIDAYS                                 0.18
## totDam8th                               0.22
## damagesDividedBy1000                    1.00
```



When we look at the pairs plot and observe the correlation between `TOTDAM` and our quantitative predictors (`DEMANDED` and `TRIDAYS`), we see that there is a rather weak positive correlation, at 0.35 for each predictor.

We can also check if there's a significant interaction between our categorical factor (`CLAIMTYPE`) and each of the quantitative predictors (`DEMANDED` and `TRIDAYS`):


Interaction Between Damages Demanded and Claim Type?
Total Damages^ (1/8) ($)
5
4
3
2
0.0e+00
2.5e+06
5.0e+06
7.5e+06
1.0e+07
Total Damages Demanded
CLAIMTYPE
Motor
Premises
Malpractice
Fraud
Rental
Other

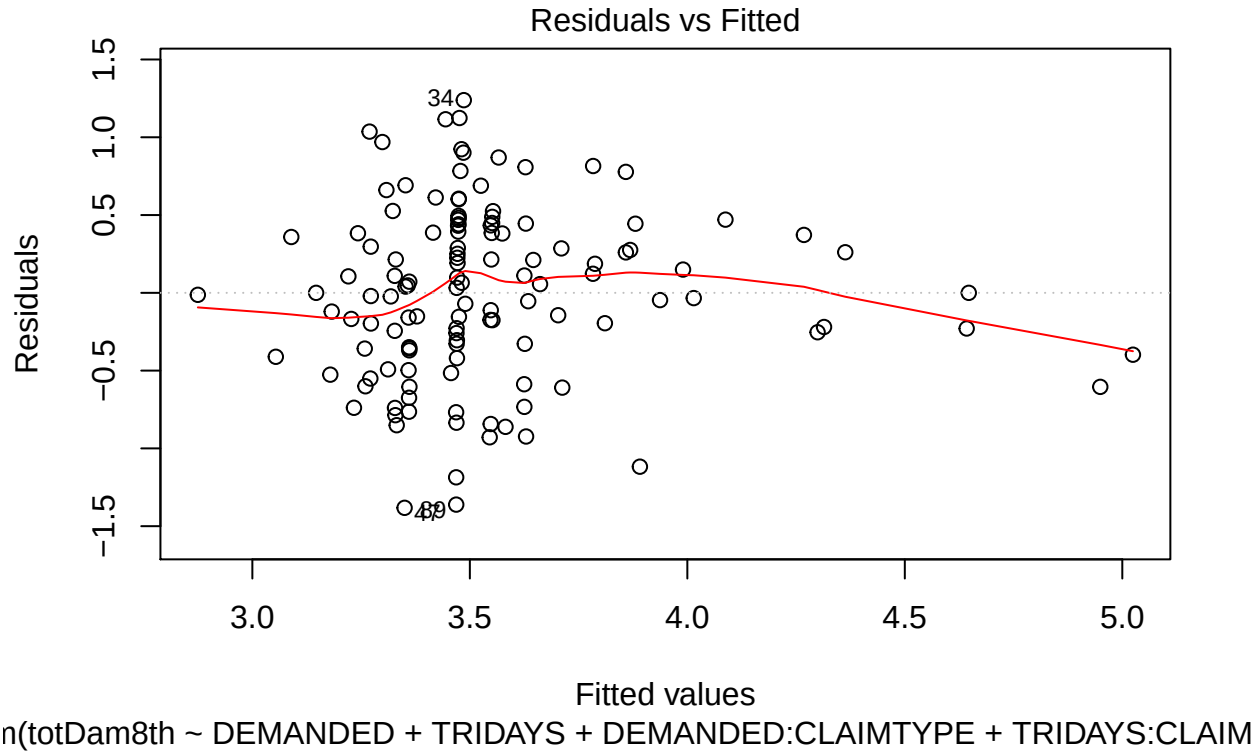


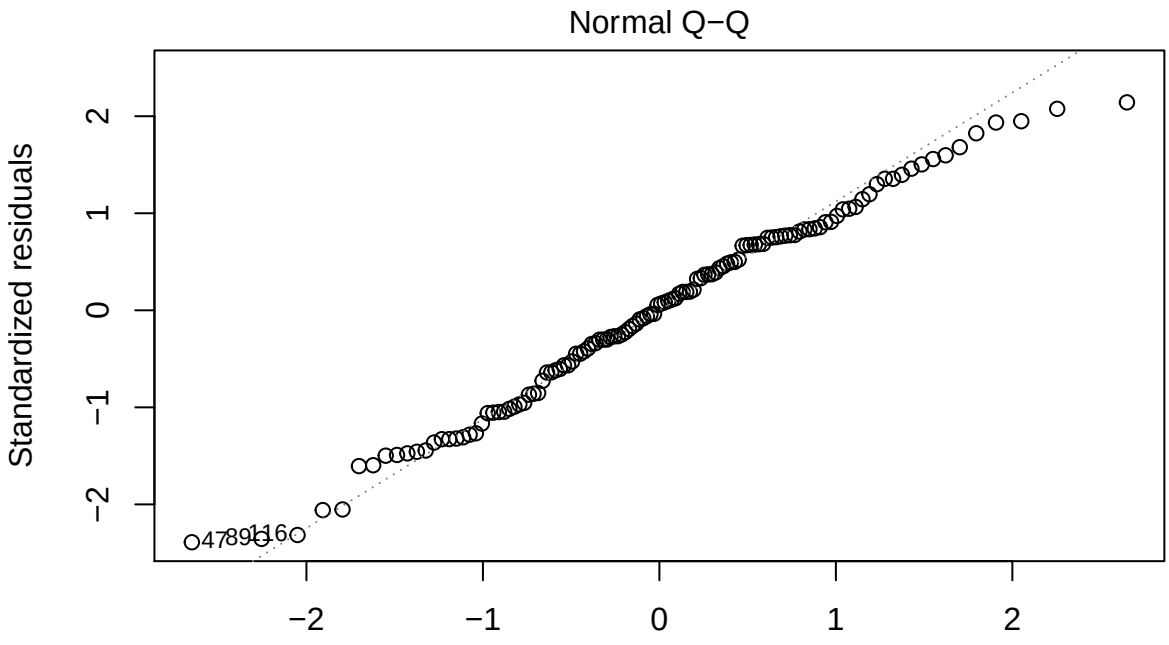
At first glance, it seems as though there is not a significant interaction between damages demanded and claim type nor is there a significant interaction between length of trial and claim type. However, we should run an interaction model just to be safe...

```
##
## Call:
## lm(formula = totDam8th ~ DEMANDED + TRIDAYS + DEMANDED:CLAIMTYPE +
##     TRIDAYS:CLAIMTYPE, data = courtData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38200 -0.35827  0.01603  0.39193  1.23856
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.391e+00  7.786e-02  43.550  < 2e-16 ***
## DEMANDED                      -9.146e-08  1.174e-07  -0.779  0.43757
## TRIDAYS                       -2.945e-02  3.687e-02  -0.799  0.42615
## DEMANDED:CLAIMTYPEPremises     3.397e-07  1.657e-07   2.050  0.04268 *
## DEMANDED:CLAIMTYPEMalpractice  8.404e-06  7.720e-06   1.089  0.27868
## DEMANDED:CLAIMTYPEFraud        2.501e-07  2.002e-07   1.250  0.21399
## DEMANDED:CLAIMTYPERental       1.534e-05  1.003e-05   1.529  0.12914
## DEMANDED:CLAIMTYPEOther        1.959e-07  1.639e-07   1.196  0.23438
## TRIDAYS:CLAIMTYPEPremises     -1.403e-01  1.741e-01  -0.806  0.42215
## TRIDAYS:CLAIMTYPEMalpractice  -2.979e-01  6.594e-01  -0.452  0.65225
## TRIDAYS:CLAIMTYPEFraud         5.102e-02  7.874e-02   0.648  0.51831
## TRIDAYS:CLAIMTYPERental       -2.519e-01  2.758e-01  -0.913  0.36307
```

```
## TRIDAYS:CLAIMTYPEOther         1.067e-01  3.644e-02   2.929  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.582 on 113 degrees of freedom
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.1937
## F-statistic: 3.502 on 12 and 113 DF,  p-value: 0.0002034
```

From the regression output, we see that there both interaction terms are significant and thus we must retain all the explanatory predictors (and the respective vifs do not matter because we have a significant interaction). Now, we move on to the residual diagnostics to determine if the assumptions for our errors are met.





From the residual diagnostics, we see that the errors are reasonably independent because the residuals appear to be generally scattered above and below the zero line. We can also state that the mean is 0 because the residuals are reasonably centered around 0. Additionally, there appears to be constant spread above and below the zero line. From the qq plot, we gather that points are generally close to the diagonal, aside from the data points at the positive extremes which is enough for us to satisfy the assumption for normality.

```
##               GVIF Df GVIF^(1/(2*Df))
## DEMANDED  1.271082  1        1.127423
## TRIDAYS   1.057519  1        1.028358
## CLAIMTYPE 1.244309  5        1.022099
```

Even though we don't need to look for vifs because we have a significant interaction, if we calculate the vifs for a multiple linear regression we see that none of them are above 2.5 which means we don't have any multicollinearity issues either way.

Here are the summary statistics for our interaction model again:

```
##
## Call:
## lm(formula = totDam8th ~ DEMANDED + TRIDAYS + DEMANDED:CLAIMTYPE +
##     TRIDAYS:CLAIMTYPE, data = courtData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38200 -0.35827  0.01603  0.39193  1.23856
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    3.391e+00  7.786e-02  43.550  < 2e-16 ***
## DEMANDED                      -9.146e-08  1.174e-07  -0.779  0.43757
## TRIDAYS                       -2.945e-02  3.687e-02  -0.799  0.42615
## DEMANDED:CLAIMTYPEPremises     3.397e-07  1.657e-07   2.050  0.04268 *
## DEMANDED:CLAIMTYPEMalpractice  8.404e-06  7.720e-06   1.089  0.27868
## DEMANDED:CLAIMTYPEFraud        2.501e-07  2.002e-07   1.250  0.21399
## DEMANDED:CLAIMTYPERental       1.534e-05  1.003e-05   1.529  0.12914
## DEMANDED:CLAIMTYPEOther        1.959e-07  1.639e-07   1.196  0.23438
## TRIDAYS:CLAIMTYPEPremises     -1.403e-01  1.741e-01  -0.806  0.42215
## TRIDAYS:CLAIMTYPEMalpractice  -2.979e-01  6.594e-01  -0.452  0.65225
## TRIDAYS:CLAIMTYPEFraud         5.102e-02  7.874e-02   0.648  0.51831
## TRIDAYS:CLAIMTYPERental       -2.519e-01  2.758e-01  -0.913  0.36307
## TRIDAYS:CLAIMTYPEOther         1.067e-01  3.644e-02   2.929  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.582 on 113 degrees of freedom
## Multiple R-squared:  0.2711, Adjusted R-squared:  0.1937
## F-statistic: 3.502 on 12 and 113 DF,  p-value: 0.0002034
```

Although the multiple regression model may have come up with a higher coefficient of determination, since we tested whether there was an interaction and found actually two significant interactions, we can reasonably use this model to predict the total damages(^1/8) awarded to the plaintiff. We know this model is significant because of the p-value of 0.0002034. The beta1 and beta2 coefficients are both negative which tell us that as plaintiffs demand more money or as the trial goes on for a longer period of time, the total damages(^1/8) awarded to the plaintiff decreases.

Prediction

Now that we have a model that reasonably satisfies all assumptions, we are interested in predicting the amount awarded to a plaintiff who demands $100,000, has a trial of five days long, and a malpractice claimtype.

The predicted damages awarded (to the 1/8th power) is as follows:

```
3.391 + (-9.146 * 10^-8) * 100000 + (-2.945 * 10^-2)*5 +
  (8.404* 10^-06)*1 + (-2.979 * 10^-01)*1
```

```
## [1] 2.936712
```

Note: The software-created dummy variables `DEMANDED:CLAIMTYPEMalpractice` and `TRIDAYS:CLAIMTYPEMalpractice` are understood to be 1 for Malpractice claims and 0 for any other claim types.

Therefore, we predict that the damages awarded (to the 1/8th power) for plaintiff who demands $100,000 with a trial of five days long and a malpractice claim are 2.936712.

Discussion

In this analysis, we learned that the total damages awarded (to the 1/8th power) to plaintiffs can be modeled by the total damages demanded by the plaintiff, the length of the trial, and the type of claim made through an interaction model.

Initially, the distribution of total damages awarded (the response variable) was clearly skewed right and so we needed to transform this variable by raising it to the 1/8th power which made the resultant distribution reasonably symmetric. Nevertheless, this made intepreting our results much more difficult which is a limitation

of this analysis. However, the outliers in the damages demanded predictor variable are worth looking into, as they may have affected our model.

It is interesting, however, that we can say the total damages awarded (raised to the 1/8th power) to the plaintiffs are significantly impacted by the type of claim they make. Perhaps this raises some additional questions about the influence of certain factors on the amount of damages awarded to plaintiffs in civil cases.