Predicting Movie Popularity on IMDb

Nathan Dennis, Maxwell Wang, David Sharkansky

11/15/2022

Section 1: Introduction

Description: Our data set is from Kaggle and the specific data was extracted from imdb.com. IMDb, which stands for "Internet Movie Database," is an online media platform that compiles information about movies and shows in addition to allowing anyone to leave reviews about video content. Our data set represents the IMDb rating of various movies and tv series as of 2022; for this project we will focus on movies. These films are rated on an IMDb scale 0 through 10, 0 meaning the movie is "bad" and 10 meaning the movie was "good". Many factors are provided in the data set, such as number of votes, year produced, level of violence, level of nudity, etc, which may influence a film's rating. Many people use a film's IMDb rating to judge how successful or good a movie is.

Reason for choosing dataset: We were all interested in movies and what factors may affect how many votes, or ratings, a movie receives. This data set had many relevant variables that we were interested in, since many of them may affect how many votes a movie receives. One key detail that we loved about this data set was the level of specific categories, such as violence and nudity. For instance, we thought that movies with higher levels of violence may receive less votes since those films may be R-rated for adults only, meaning children and younger adults may not watch and rate the movie, resulting in fewer votes in IMDb overall. Levels are measured by None, Mild, Moderate, and Severe, ranging from the least amount to greatest amount of a variable.

Population/Sample: The population in this data set is the number of movies and films within the IMDb database. The year of the productions range from 1922 to 2022, which is over the past 100 years.

Cross-Sectional or Longitudinal study: The IMDb ratings of movies can change over time, as the number of votes from individuals increases. This means this is a longitudinal study, since the observation of IMDb rating can change over time depending on new votes.

Software: We will be using RStudio version 2022.07.2+576, with R Version 4.2.2.

Level of significance: We plan to use a 0.05 level of significance.

Missing Values: We do have some missing values in the data set. For the variable Rate, which represents the IMDb rating, there are many values with "No Rate", we will simply remove those values. In turn, the descriptive categorical variables: Nudity, Violence, Alcohol, Profanity, and Violence all have various instances with "No Rate" as the rating; these will also be removed. Finally, those with a Duration of 'None' will also be removed.

Importance: This topic is important since it could be helpful for film companies to figure out what factors may influence both the IMDb rating and number of votes a movie received. We know that the number of reviews left by viewers of a movie correlates well with the number of people who watched a movie (Lee et. al., 2014). This means that by studying what factors affect the number of votes left on IMDb, we gain insight into what factors will affect the overall popularity of a movie (how many people have watched it).

Furthermore, if someone is searching for a movie to watch, they may see what movies have a higher IMDb rating and high number of votes. According to (Pentheny, 2015), reading positive reviews of movies influences a customer's decision to buy movie tickets. Rating alone is not enough to justify watching a movie, since many consumers may be seeking to watch movies that have been rated by many people, as the number of ratings generally describes its trustworthiness.

Research Questions:
1: **Is Profanity in a movie associated with Violence?**
For this first question we will use a chi-squared test with the categorical variables Violence and Profanity. The Violence variable has four levels in order, None, Mild, Moderate, and Severe, assessing the level of Violence within a film. The Profanity variable has four levels in order, None, Mild, Moderate, and Severe, assessing the level of Profanity within a film. We will validate all assumptions for the chi-squared test before we conduct it. Once we conduct a chi-squared test we can decide to run a post-hoc analysis if we find out that the two variables are associated, using the Violence level of None. We want to observe if there is any association between these two variables since these variables may influence the popularity of the production and influence each other. This could also help us decide whether we include one or both of these predictors in our regression model.

2: **How does the frightening variable affect the number of votes for a movie?**
For this question we plan to validate assumptions and then use an ANOVA test to analyze how much the frightening variable affects the total number of votes a movie receives. We will need to use the frightening column as well as the votes column from the dataset. Frightening is our independent and categorical variable with 4 levels, None, Mild, Moderate, and Severe, measuring the level of Frightening in a production. Votes, representing the number of votes a production receives, is our dependent and continuous numerical variable. We would like to see if the various frightening levels have a significant effect on the number of votes a movie receives on IMDb. If we find that the number of votes is not normally distributed for each level, we will use a transformation on the variable. If we find that there is a significant difference in means in at least one level of Frightening, we can conduct a Tukey's test to see which groups have a significantly different mean number of votes.

3: **Which factors significantly predict the number of votes that a movie gets?**
We will perform multiple linear regression and create three models to examine potential predictor variables for the number of votes that a movie gets. Then we will compare the models, so we can narrow down the most important predictors and formulate a simpler model. We will need the Votes variable as well as Rate, Duration, and the "content" variables (Nudity, Violence, Alcohol, Frightening, and Profanity). By building a regression model, we can see which variables most significantly influence the number of votes a movie gets to directly answer, because as discussed in the Importance section, more votes correlates with more ticket sales, which is highly relevant for the movie industry.

Section 2:

Variables and Description:

Variable Name	Type of Data	Level	Units
Name	Categorical	Ordinal	Movie Name
Date	Numerical	Interval	Year
Rate	Numerical	Interval	Points
Votes	Numerical	Continuous	Number of Votes
Genre	Categorical	Nominal	Type of Genre
Duration	Numerical	Continuous	Minutes
Type	Categorical	Nominal	Series or Film
Certificate	Categorical	Nominal	Type of Certificate
Nudity	Categorical	Ordinal	4 Levels*
Violence	Categorical	Ordinal	4 Levels*
Profanity	Categorical	Ordinal	4 Levels*
Alcohol	Categorical	Ordinal	4 Levels*
Frightening	Categorical	Ordinal	4 Levels*

*Levels include, in order, None, Mild, Moderate, Severe.

Visualizations:

```r
movies$Profanity <- factor(movies$Profanity, levels = c("None",
    "Mild", "Moderate", "Severe"))

movies$Violence <- factor(movies$Violence, levels = c("None",
    "Mild", "Moderate", "Severe"))

ggplot(movies, aes(x = Profanity, fill = Violence)) + geom_bar(position = "fill") +
    labs(title = "Amount of Profanity in Films compared to Violence",
        subtitle = "Movies from 1922 to 2022", y = "Proportion",
        x = "Amount of Profanity", fill = "Amount of Violence")
```

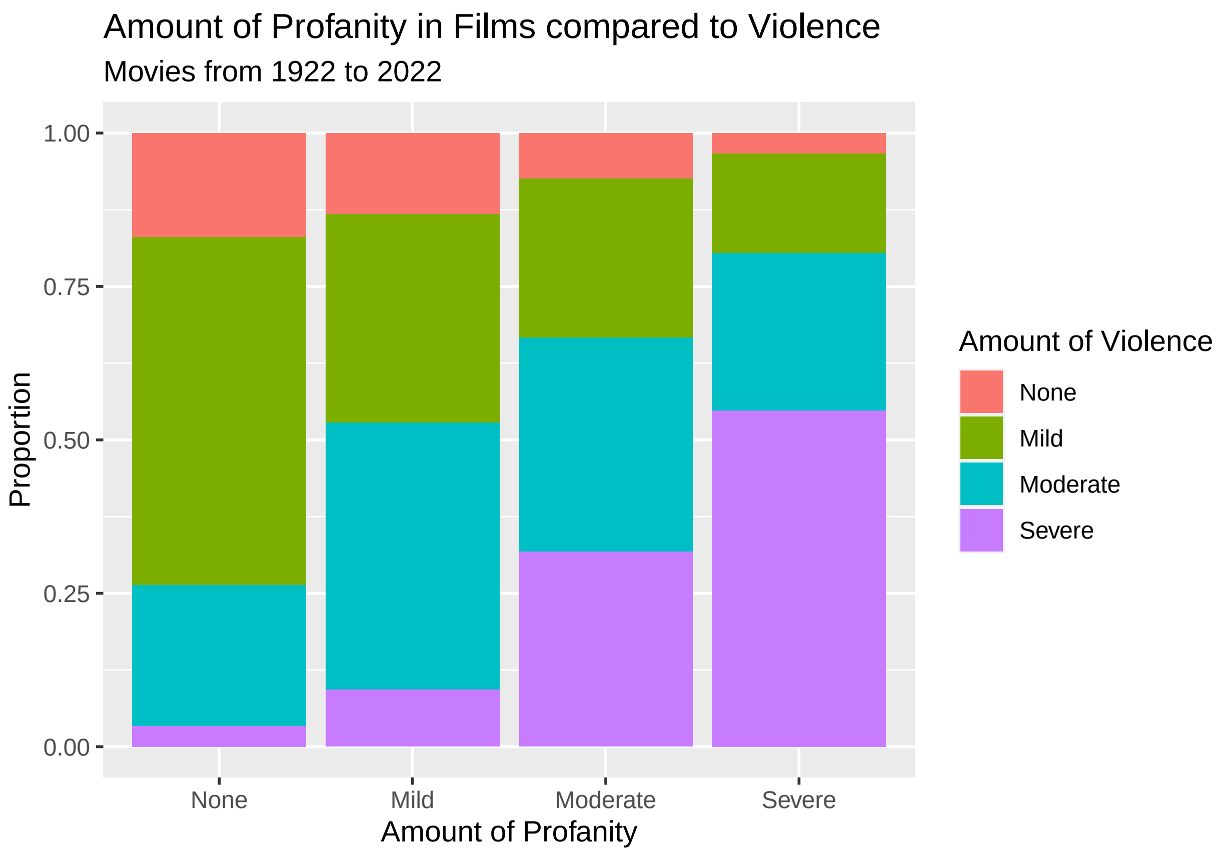


Figure 1

According to the above mosaic plot, we can see that movies with none or mild profanity tend to have mostly mild or moderate violent content, whereas movies with moderate to severe profanity have more and a majority of moderate to severe violence. This visualization supports the claim that profanity is associated with violence in movies because as the amount of profanity increases from none to mild to moderate to severe, we see that the proportion of movies with no violence whatsoever or mild violence decreases with every increase in profanity, while the proportion of movies with severe violence increase substantially. As indicated in the purple color on the graph, representing the proportion of films with a severe level of violence, the proportion of a severe level of violence increases as amount of profanity increases.

```
movies$Frightening <- factor(movies$Frightening, levels = c("None",
    "Mild", "Moderate", "Severe"))

ggplot(movies, aes(x = Frightening, y = Votes, fill = Frightening)) +
    geom_boxplot() + labs(title = "Vote count by Frightening Level",
    x = "Frightening Level", y = "Number of Votes", fill = "Frightening Level") +
    scale_fill_viridis_d()
```

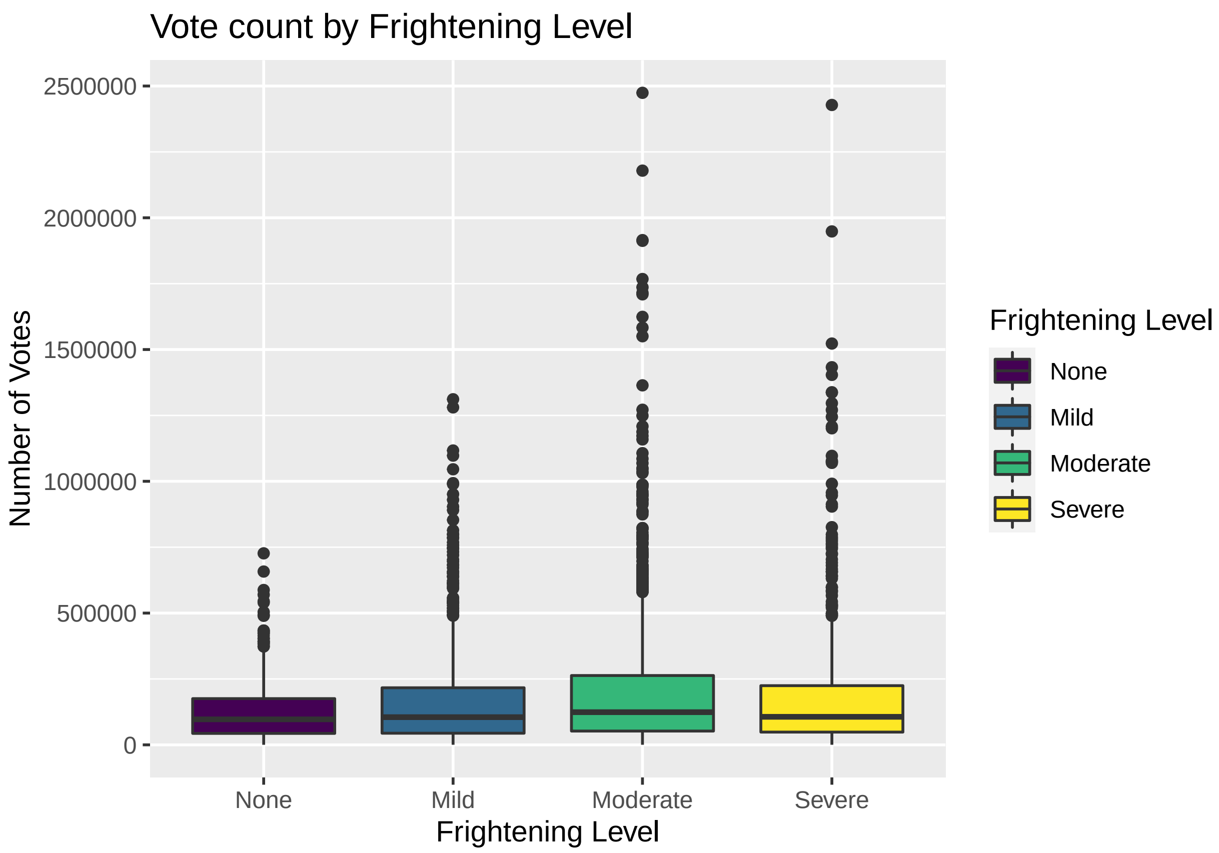


Figure 2-1

We can see in the boxplot above that the Vote count for different levels of Frightening within movies are very scattered. The median seems to be equal for each level, with a smaller spread for films with a 'none' frightening rating. The IQR is relatively small, but there are many outliers in all four of the frightening levels. The outliers are concerning, as we see there are many significant ones that may affect our analysis. We can try a log transformation to see if we can make the votes normally distributed for each level of Frightening.

```
ggplot(movies, aes(x = Frightening, y = log(Votes), fill = Frightening)) +
    geom_boxplot() + labs(title = "Log Vote count by Frightening Level",
    x = "Frightening Level", y = "Log Number of Votes", fill = "Frightening Level") +
    scale_fill_viridis_d()
```

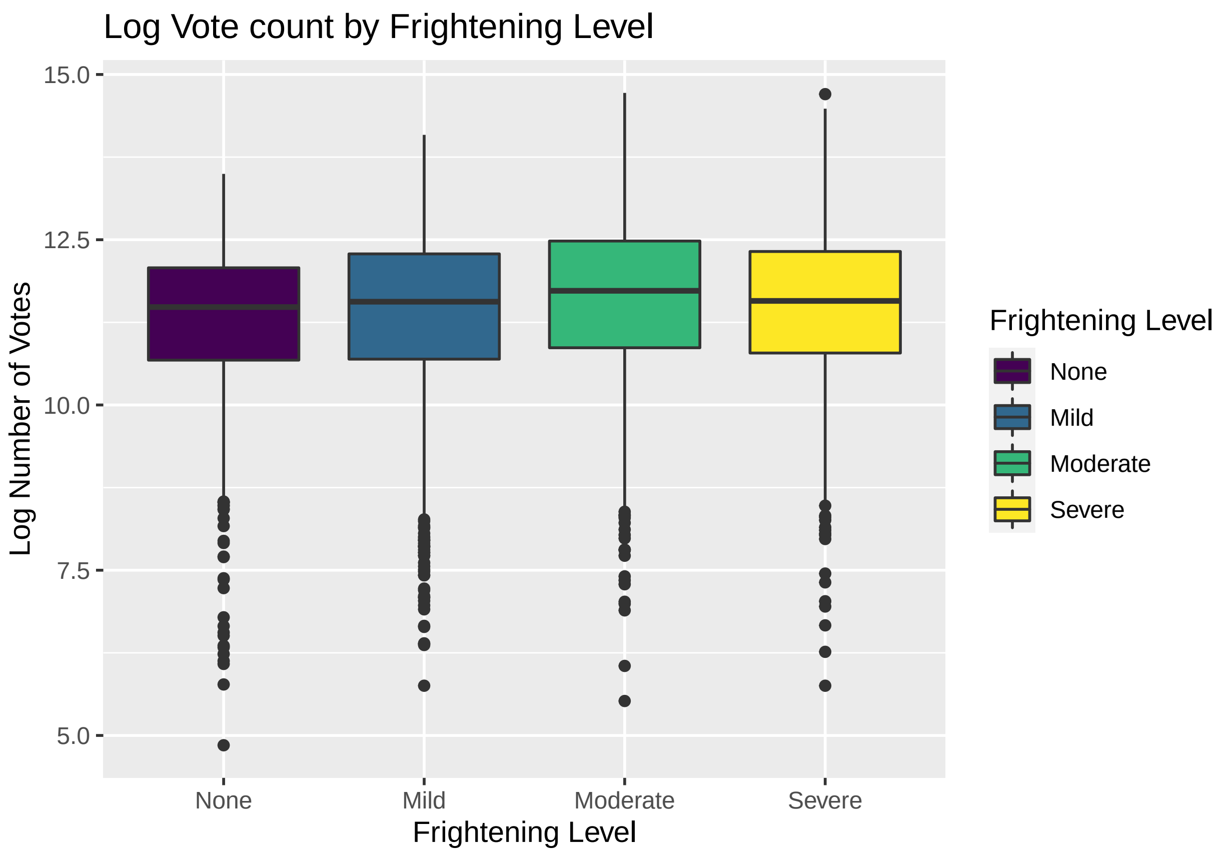


Figure 2-2

We decided to use a logarithmic transformation to better normalize the dependent variable. From the boxplot above, we can see that apart from a slight left skew with some outliers, our median sits in between the first and third quartiles for all categories at around the same point, approx 11.5. The spread of each variable is also approximately the same, as the IQR is not much different in each boxplot with some outliers towards the bottom of the plot for each level. We can conclude that means log(Votes) is approximately normal for each level of the frightening variable.

```r
ggplot(movies, aes(x = Rate, y = Votes, col = Violence)) + geom_point() +
    labs(title = "Votes vs IMDb Rating For Films", subtitle = "Colored by Level of Violence",
        x = "IMDb Rating", y = "Number of Votes")
```

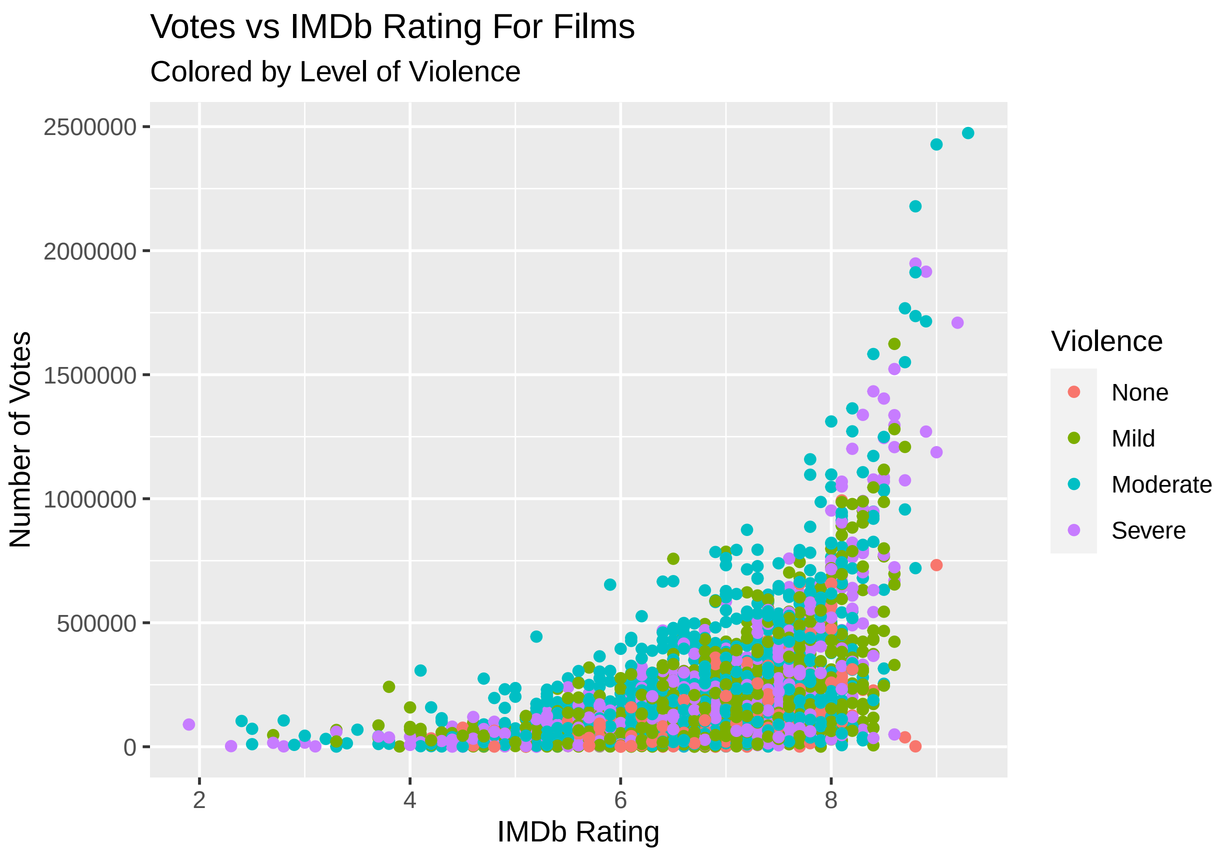


Figure 3-1

We can observe in the scatterplot between Number of Votes (Votes) vs IMDb Rating (Rate) colored by level of Violence, the variables seem to be related to each other approximately linearly. The variables experience a weak, positive correlation, as the points slightly increase with each other. As Rate increases, the number of votes also slightly does as indicated in the graph. There don't seem to be any significant outliers. When we color the data by Violence we see the effect that violence has on the Number of Votes. The most popular movies with the highest number of votes and highest IMDb rating are mostly ones with Moderate or Severe levels of Violence.

```
ggplot(movies, aes(x = Date, y = Votes, col = Violence)) + geom_point() +
    labs(title = "Votes vs Date of Film Production", subtitle = "Colored by Level of Violence",
        x = "Date of Production", y = "Number of Votes")
```

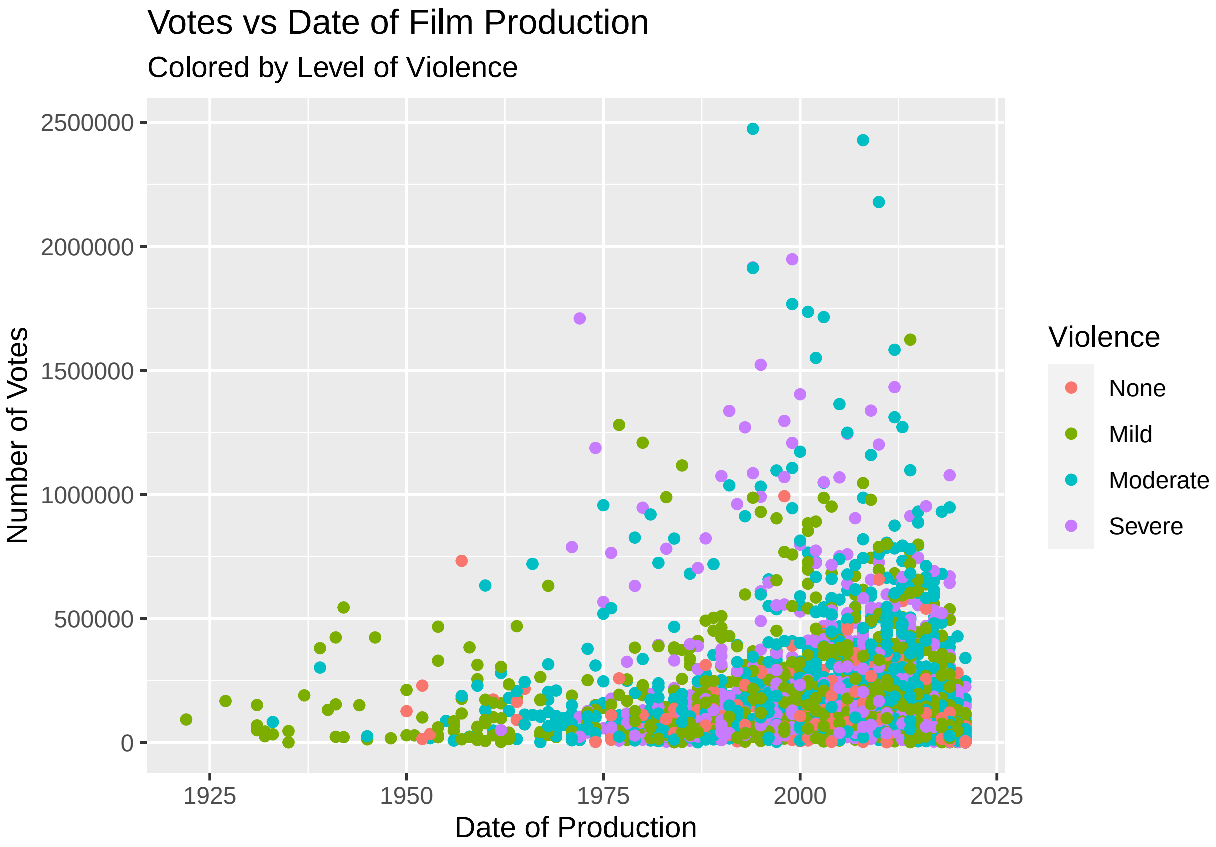


Figure 3-2
We can observe in the scatterplot between Date (Year) vs Number of Votes (Votes) colored by level of Violence, the variables seem to be related to each other approximately linearly. The data experiences a weak, positive correlation, as the points slightly increase with each other. As the date a film was created in increases, the number of votes also does. There are some significant outliers where there are a lot of votes. Movies that were released closer to 2022 (modern day) have seen a steady increase in the number of votes received. There does not seem to be any obvious relationship in this graph with the level of Violence.

Figure 4-1 : Table for Violence and Profanity

	Profanity			
Violence	None	Mild	Moderate	Severe
None	65	154	78	21
Mild	217	395	270	103
Moderate	88	507	365	162
Severe	13	108	332	347

An important observation is that every cell has a count greater than 5, which is necessary for a chi-squared test. We don't see many obvious patterns between the variables, the count seems to be somewhat spread through each cell. It is noticeable though that the 'None' level of both variables seems to consistently have the smallest frequency in their cells.

Section 3: Statistical Results

Question 1

Is Profanity in a movie associated with Violence?

To figure out if Profanity is associated with Violence within films we will run a chi-squared test to compare proportions. To run a chi-squared test, we must meet all of the assumptions. First, both the categorical variables are measured at an ordinal level, as displayed in the variables description. We also know that both variables consist of 4 categorical, independent groups which are None, Mild, Moderate, and Severe, and are independent since for example no film can have more than one level of violence or profanity. Then, we observe that the sample size for every cell have counts greater than 5 in figure 4-1. We can then proceed with the test.

We are measuring one statistic, which will be the chi squared statistic. There are two variables with 4 categories each, so we will have (4 - 1) * (4 - 1) = 9 degrees of freedom.

Null hypothesis H_0: There is no association between Violence and Profanity in movies.
Alternative hypothesis H_a: There is an association between Violence and Profanity in movies.

```
chisq.test(movies$Violence, movies$Profanity)
```

```
##
##  Pearson's Chi-squared test
##
## data:  movies$Violence and movies$Profanity
## X-squared = 696.98, df = 9, p-value < 2.2e-16
```

From our chi-squared test, we observe $X^2 = 696.98$ and our degrees of freedom of 9, which was also calculated above. Because the associated p-value of $2.2e^{-16}$ is less than our significance level of 0.05, $2.2e^{-16} < \alpha = 0.05$, we have sufficient evidence to reject the null hypothesis. Therefore, we conclude that there is an association between Violence and Profanity in movies.

We can do a post-hoc test to determine whether the specific level for Violence, None, has a statistically significant difference in the number of votes for each level of Profanity. We are interested in this since we want to observe how the level None for Violence relates to the proportion of each level of Profanity. We also observed in the table that the Violence level None consistently had very smaller counts, which is interesting and something we wanted to analyze. We will run a difference of proportions.

```
ViolenceNone <- c(65, 154, 78, 21)

Total <- c(383, 1164, 1045, 633)

prop.test(ViolenceNone, Total)
```

```
##
##  4-sample test for equality of proportions without continuity correction
##
## data:  ViolenceNone out of Total
## X-squared = 73.899, df = 3, p-value = 6.237e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4
## 0.16971279 0.13230241 0.07464115 0.03317536
```

With a p-value of nearly 0, 6.237e-16, which is below our significance level of 0.05, we can conclude that there is a statistically significant difference in number of votes between the Violence levels None and the total number of votes for each level of Profanity. We can proceed with out difference in proportions post hoc test below.

```r
p = c(0.16971279, 0.13230241, 0.07464115, 0.03317536)
N = length(p)
k = N - 1
NN = 3225  #Obtained from adding total observations in each group (Total vector)`
value = critical.range = c()


for (i in 1:(N - 1)) {
    for (j in (i + 1):N) {
        value = c(value, (abs(p[i] - p[j])))
        critical.range = c(critical.range, sqrt(qchisq(0.95,
            k)) * sqrt(p[i] * (1 - p[i])/NN + p[j] * (1 - p[j])/NN))
    }
}


round(cbind(value, critical.range), 3)
```

```
##      value critical.range
## [1,] 0.037          0.025
## [2,] 0.095          0.023
## [3,] 0.137          0.020
## [4,] 0.058          0.021
## [5,] 0.099          0.019
## [6,] 0.041          0.016
```

For the first comparison, between the first and second proportion which represents the proportions of Violence being None and Mild, we have that the critical value of 0.025 is less than the value 0.037, so the difference in proportions is statistically significant between the two groups.

For the second comparison, between the first and third proportion which represents the proportions of Violence being None and Moderate, we have that the critical value of 0.023 is less than the value 0.095, so the difference in proportions is statistically significant between the two groups.

For the third comparison, between the first and fourth proportion which represents the proportions of Violence being None and Severe, we have that the critical value of 0.020 is less than the value 0.137, so the difference in proportions is statistically significant between the two groups.

For the fourth comparison, between the second and third proportion which represents the proportions of Violence being Mild and Moderate, we have that the critical value of 0.021 is less than the value 0.058, so the difference in proportions is statistically significant between the two groups.

For the fifth comparison, between the second and fourth proportions which represent the proportions of Violence being Mild and Severe, we have that the critical value of 0.019 is less than the value 0.099, so the difference in proportions is statistically significant between the two groups.

For the final comparison, between the third and fourth proportion which represents the proportions of Violence being Moderate and Severe, we have that the critical value of 0.016 is less than the value 0.041, so the difference in proportions is statistically significant between the two groups.

We observe that between each group the differences are statistically significant.

Question 2

How does the frightening variable affect the number of votes for a movie?

To figure out how the Frightening variable affects the number of votes for a movie we will conduct an ANOVA test to determine what levels of Frightening have an affect on the number of Votes a film receives. We will be using the transformation log on the number of votes since this makes the data normally distributed for each level of frightening as shown in figure 2-2.
Assumptions:
We can verify all the assumptions for running the ANOVA test. We first will use a single dependent variable, log Votes, and independent variable, Frightening which is categorical with 4 total groups, None, Mild, Moderate, and Severe, which are independent of each other since no film can have more than 1 level of Frightening. Based on the Boxplot in Figure 2-1, we first see that with Votes vs Frightening, the distribution is heavily skewed left for each value and is not normally distributed for each level of Frightening. We applied a log transformation to the dependent variable (Figure 2-2) to approximate normality. In the new figure, note that although the distribution of number of votes for each level of Frightening is slightly left skewed, the median is pretty evenly distributed between the first and third quartile, so we will conclude that log(Votes) is approximately normally distributed for all levels of Frightening. There were some outliers in the boxplot, but we concluded they were not extremely significant. The next assumption is that the dependent variable is continuous, and since log(Votes) is continuous this holds. Every movie corresponds to exactly one value in the dependent variable, the log number votes that movie received. Each participant, person in this case, only casts one vote per film. Again, within the boxplot we also concluded that there were no significant outliers and the assumption was met. We can test for the variance assumption below:

```
leveneTest(log(movies$Votes) ~ movies$Frightening)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    3  1.4566 0.2244
##       3221
```

We also see that the variances are equal for each level of Frightening after running Levene’s test, with a p-value of 0.2244, greater than our significance level of 0.05. Since we have met all assumptions required to conduct an ANOVA test, we can proceed.

Null Hypothesis H_0: There is no difference in the true mean of number of log Votes with respect to any of the four Frightening levels for movies rated on IMDb.
Alternative Hypothesis H_a: There is a difference in the true mean of number of log Votes with respect to at least one of the four Frightening levels for movies rated on IMDb.

```
modelq2 <- aov(log(movies$Votes) ~ movies$Frightening)
summary(modelq2)
```

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## movies$Frightening    3     77  25.701   14.29 3.01e-09 ***
## Residuals           3221   5794   1.799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test obtains a p-value of 3.01e-09 which is less than the significance level of $\alpha = 0.05$. Therefore, we reject the null hypothesis and conclude that there is a difference in the true mean of the number of log votes with respect to one of the four Frightening levels.

To find out specific differences in means between levels, we chose to use a Tukey's HSD post-hoc test in order to reveal if there is a statistically significant difference between all each group of Frightening levels. We wanted to compare each level of Frightening to one another with both p-values and 95% confidence intervals. We observed that the sample sizes were significantly different between each level of Frightening. More specifically, the counts for each level of Frightening were 430 for None, 925 for Mild, 1252 for Moderate, and 618 for Severe. Tukey's test is known to be the most useful test when obtaining confidence intervals for the difference between means when the sample sizes are not equal. Furthermore, Tukey's test is also useful when we want to keep the level of the Type I error equal to the chosen alpha level, which we wanted to keep as low as possible (Abdi and Williams, 2010).

```
TukeyHSD(modelq2, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(movies$Votes) ~ movies$Frightening)
##
## $`movies$Frightening`
##                       diff         lwr        upr     p adj
## Mild-None        0.1405047 -0.06070923 0.34171855 0.2759359
## Moderate-None    0.4233548  0.23065997 0.61604957 0.0000001
## Severe-None      0.3086811  0.09218721 0.52517492 0.0014346
## Moderate-Mild    0.2828501  0.13338164 0.43231858 0.0000072
## Severe-Mild      0.1681764 -0.01093006 0.34728287 0.0747261
## Severe-Moderate -0.1146737 -0.28415345 0.05480604 0.3034967
```

We can see with the Tukey's test what levels of Frightening compared to each other relative to their mean log votes. We will use a significance level of 0.05.

We can see comparisons for each level of Frightening now. We will analyze these groups one by one. For all observations we will use the same null hypothesis, that there is no statistically significant difference in mean number of log votes between the two Frightening levels, and alternative hypothesis of there exists a statistically significant difference in mean number of log votes between the two Frightening levels.

First, we observe the Mild-None comparison, comparing the two levels of Frightening in films. We obtain a p-value of 0.2759359, which is greater than our significance level of 0.05, so we fail to reject the null hypothesis. We can conclude there is no statistically significant difference in mean number of log votes between the Frightening levels Mild and None.

Next, we observe the Moderate-None comparison, comparing the two levels of Frightening in films. We obtain a p-value of nearly 0, specifically 0.0000001, which is less than our significance level of 0.05, so we are able to reject the null hypothesis. We can conclude there is a statistically significant difference in mean number of log votes between the Frightening levels Moderate and None. We then can observe the 95% confidence interval, which contains only positive values ranging from 0.23065997 to 0.61604957. Since this interval contains positive values and the level Moderate is entered first into the equation, Moderate minus None, we conclude that films with a Frightening level of Moderate have a greater mean number of log votes than films with a Frightening level of None. Now we can observe the actual calculated difference in means between the groups, which was 0.4233548, so we expect the mean log number of votes for films with a Frighting level of Moderate to be 0.4233548 greater than the mean log number of votes for films with a Frightening level None.

Next, we observe the Severe-None comparison, comparing the two levels of Frightening in films. We obtain a p-value of 0.0014346, which is less than our significance level of 0.05, so we are able to reject the null hypothesis. We can conclude there is a statistically significant difference in mean number of log votes between the Frightening levels Severe and None. We then can observe the 95% confidence interval, which contains only positive values ranging from 0.09218721 to 0.52517492. Since this interval contains positive values and the level Severe is entered first into the equation, Severe minus None, we conclude that films with a Frightening level of Severe have a greater mean number of log votes than films with a Frightening level of None. Now we can observe the actual calculated difference in means between the groups, which was 0.3086811, so we expect the mean log number of votes for films with a Frighting level of Severe to be 0.3086811 greater than the mean log number of votes for films with a Frightening level None.

Next, we observe the Moderate-Mild comparison, comparing the two levels of Frightening in films. We obtain a p-value of nearly 0, specifically 0.0000072, which is less than our significance level of 0.05, so we are able to reject the null hypothesis. We can conclude there is a statistically significant difference in mean number of log votes between the Frightening levels Moderate and Mild. We then can observe the 95% confidence interval, which contains only positive values ranging from 0.13338164 to 0.43231858. Since this interval contains positive values and the level Moderate is entered first into the equation, Moderate minus Mild, we conclude that films with a Frightening level of Moderate have a greater mean number of log votes than films with a Frightening level of Mild. Now we can observe the actual calculated difference in means between the groups, which was 0.2828501, so we expect the mean log number of votes for films with a Frighting level of Moderate to be 0.2828501 greater than the mean log number of votes for films with a Frightening level Mild.

Next, we observe the Severe-Mild comparison, comparing the two levels of Frightening in films. We obtain a p-value of 0.0747261, which is slightly greater than our significance level of 0.05, so we fail to reject the null hypothesis. We can conclude there is no statistically significant difference in mean number of log votes between the Frightening levels Severe and Mild.

Finally, we observe the Severe-Moderate comparison, comparing the two levels of Frightening in films. We obtain a p-value of 0.3034967, which is greater than our significance level of 0.05, so we fail to reject the null hypothesis. We can conclude there is no statistically significant difference in mean number of log votes between the Frightening levels Severe and Moderate.

We observe that the comparisons with a statistically significant difference in mean number of log votes were Moderate-None, Severe-None, and Moderate-Mild. From this test and the results an interesting observation is the Moderate level consistently had a greater mean number of log Votes when compared to the other variables, specifically Mild and None were statistically significant with Moderate having a greater mean number of log votes. However, when compared to a Severe level of Frightening the difference was not statistically significant.

Question 3

Which factors significantly predict the number of votes that a movie gets?
We used multiple linear regression to determine how to create the best model with many different variables that influence the number of votes received by a movie.
Initial Assumptions: We begin by testing for normality. If the assumption is not met, we will try multiple transformations.

```
shapiro.test(movies$Votes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  movies$Votes
## W = 0.67796, p-value < 2.2e-16
```

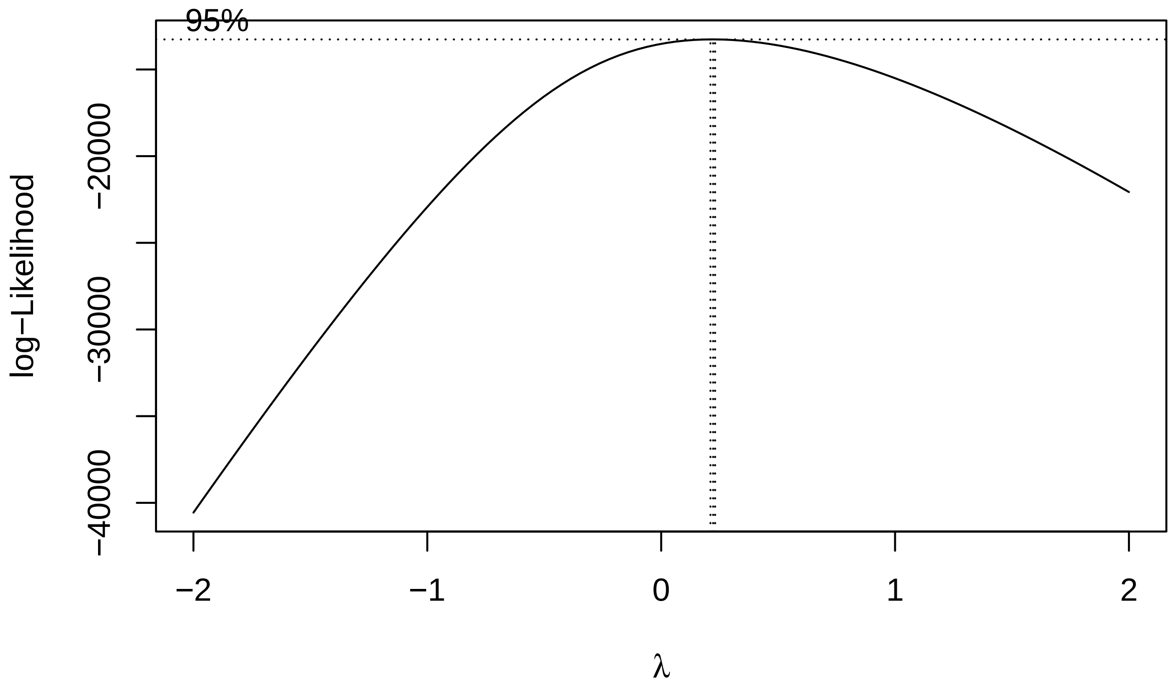
```
shapiro.test(sqrt(movies$Votes))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(movies$Votes)
## W = 0.93008, p-value < 2.2e-16
```

```
shapiro.test(log(movies$Votes))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(movies$Votes)
## W = 0.9575, p-value < 2.2e-16
```

```
bc <- boxcox(movies$Votes ~ movies$Rate)
```



```
lambda <- bc$x[which.max(bc$y)]

boxcox <- (movies$Votes^lambda - 1)/lambda
shapiro.test(boxcox)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  boxcox
## W = 0.99775, p-value = 0.0001253
```

The dependent variable, Votes (number of votes), must be approximately normally distributed, which is true if our p-value is greater than our significance level within these tests. We checked using a shapiro test, which resulted in a p-value of 2.2e-16 which is below our signifiance level of 0.05, so we conclude the Votes are not normally distributed. When using transformations such as log, square root, and boxcox, the data still wasn't normally distributed. For the log transformation and square root transformation we obtained a p-value of nearly 0, specifically 2.2e-16 which is under our significance level, meaning the data is still not normally distributed. For the boxcox test we obtained a p-value of 0.00012, less than the significance level of 0.05, meaning the data is not normally distributed. We attempted multiple boxcox transformations with different variables, yet none could make the Votes variable normal. We will assume that Votes is normally distributed. We also observe the linearity assumption and outlier assumptions within the models we created above for important numerical variables we will use in our model, where we also observed no significant outliers when comparing the variables. Visualization 3-1 showed the relationship between number of votes and IMDb rating, where we observed a weak positive correlation and no significant outliers. Since the initial assumptions are met we can proceed with creating the linear model. We will observe the final assumptions assumptions once we obtain our model.

```
base1 <- lm(Votes ~ Rate + Duration + as.factor(Nudity) + as.factor(Violence) +
    as.factor(Alcohol) + as.factor(Frightening) + as.factor(Profanity),
    data = movies)

ols_step_backward_p(base1)
```

```
## [1] "No variables have been removed from the model."
```

```
ols_step_forward_p(base1)
```

```
##
##                                        Selection Summary
## --------------------------------------------------------------------------------------------
##        Variable                              Adj.
## Step          Entered            R-Square   R-Square      C(p)         AIC           RMSE
## --------------------------------------------------------------------------------------------
##    1    Rate                       0.2459     0.2457    228.3897    87919.5388   201063.8206
##    2    Duration                   0.2751     0.2747     96.7208    87794.0794   197160.1592
##    3    as.factor(Violence)        0.2910     0.2899     26.3052    87728.8646   195086.1205
##    4    as.factor(Alcohol)         0.2937     0.2919     15.8256    87722.4316   194801.2301
##    5    as.factor(Profanity)       0.2957     0.2932      8.8147    87719.4243   194620.1804
##    6    as.factor(Nudity)          0.2979     0.2948      0.6874    87715.2711   194404.8399
##    7    as.factor(Frightening)     0.2989     0.2952     -2.0000    87716.5608   194353.7632
## --------------------------------------------------------------------------------------------
```

Our base model begins with Rate, Duration, Nudity, Violence, Alcohol, Frightening, and Profanity. We opted to not include Name (since each movie has a unique name), Genre and Certificate (because there are more than 12 categories for each of these variables with some categories having very few entries), Type (which was already used to filter for films only), and Date since the Date variables ranges from 1922 to 2022 and it wouldn't make sense to include.

When we apply the step backward method to remove variables to our base model, the function tells us to not remove any variables. We will decide what variables to remove with the step forward function and manually.

First, we remove Profanity from the model because in question 1 observe that Profanity was highly associated with Violence, so having both predictors in the model is unnecessary. We will also remove the variable Frightening, since it is the last and therefore least important predictor in our model according to the step forward function. We know that step-forward will add the most significant predictors in sequential order (STHDA, 2018), meaning the least impactful predictors will be the final steps of the model.

```
base2 <- lm(Votes ~ Rate + Duration + as.factor(Nudity) + as.factor(Violence) +
    as.factor(Alcohol), data = movies)

ols_step_forward_p(base2)
```

```
##
```

```
##                                       Selection Summary
## -----------------------------------------------------------------------------------------------
##        Variable                                 Adj.
## Step         Entered             R-Square     R-Square      C(p)          AIC             RMSE
## -----------------------------------------------------------------------------------------------
##    1     Rate                       0.2459       0.2457    218.9127    87919.5388    201063.8206
##    2     Duration                   0.2751       0.2747     87.6110    87794.0794    197160.1592
##    3     as.factor(Violence)        0.2910       0.2899     17.3944    87728.8646    195086.1205
##    4     as.factor(Alcohol)         0.2937       0.2919      6.9492    87722.4316    194801.2301
##    5     as.factor(Nudity)          0.2957       0.2932      0.0000    87719.4615    194621.3013
## -----------------------------------------------------------------------------------------------
```

Next we used the step function for this model, and again nothing was removed. We can manually observe the rankings for each predictor variable, from most to least important. The bottom 2 variables ranked by importance were Alcohol and Nudity, so we chose to remove both those variables from this model.

```r
base3 <- lm(Votes ~ Rate + Duration + as.factor(Violence), data = movies)
```

We will decide to end with these three models and will compare the models to analyze the results.

```r
compareLM(base1, base2, base3)
```

```
## $Models
##   Formula
## 1 "Votes ~ Rate + Duration + as.factor(Nudity) + as.factor(Violence) + as.factor(Alcohol) + as.facto
## 2 "Votes ~ Rate + Duration + as.factor(Nudity) + as.factor(Violence) + as.factor(Alcohol)"
## 3 "Votes ~ Rate + Duration + as.factor(Violence)"
##
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC R.squared Adj.R.sq    p.value Shapiro.W
## 1   18   3207 87720 87720 87830    0.2989   0.2952 4.778e-232    0.8243
## 2   12   3213 87720 87720 87800    0.2957   0.2932 6.337e-235    0.8253
## 3    6   3219 87730 87730 87770    0.2910   0.2899 3.432e-237    0.8264
##   Shapiro.p
## 1 1.012e-50
## 2 1.263e-50
## 3 1.606e-50
```

To compare all of the linear models, we decided to use the BIC over AIC because we were not seeing very much variance in the AIC of our models. This is likely due to BIC penalizing extra parameters even more than AIC does. This supports our research, because given our base model and factoring all the Violence, Profanity, and similar categorical variables, there are simply too many variables in the model, BIC would benefit from the removal of some of them.

From the last round of comparisons, it is evident that the model Base - Alcohol - Nudity - Frightening - Profanity has the lowest BIC which means it has the best fit when compared to the other models. This is likely due to it having the fewest extraneous variables while preserving the important variables that most strongly influence Votes. Furthermore, given that the BIC values are similar, we would naturally choose the simplest model. We will choose our third and final model, which is the simplest model, since all models

predict equally well.
Final Model:

```
final <- lm(Votes ~ Rate + Duration + as.factor(Violence), data = movies)

summary(final)
```

```
##
## Call:
## lm(formula = Votes ~ Rate + Duration + as.factor(Violence), data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -446955 -110777  -31375   66620 1940458
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -793912.9    28031.1 -28.323  < 2e-16 ***
## Rate                         104941.1     3863.5  27.162  < 2e-16 ***
## Duration                       1817.8      182.5   9.961  < 2e-16 ***
## as.factor(Violence)Mild       48577.2    12583.5   3.860 0.000115 ***
## as.factor(Violence)Moderate   93501.2    12490.4   7.486 9.12e-14 ***
## as.factor(Violence)Severe     85563.9    12984.8   6.590 5.14e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195100 on 3219 degrees of freedom
## Multiple R-squared:  0.291,  Adjusted R-squared:  0.2899
## F-statistic: 264.2 on 5 and 3219 DF,  p-value: < 2.2e-16
```

Our equation for number of votes is: $\hat{Votes} = -793912.9 + 104941.1 * Rate + 1817.8 * Duration + 48577.2 * Violence(Mild) + 93501.2 * Violence(Moderate) + 85563.9 * Violence(Severe)$.

We obtain an intercept of -793912.9, which means that when every variable is held at a constant 0, with a baseline/reference of the level of Violence being None, we expect the number of votes for a film to be -793912.9 The intercept with every variable at 0 has no true meaning since we can't have negative votes on a movie, but when all the other variables take on natural realistic values that are not 0, we will nearly always get a positive number.

We obtain a slope of 104941.1 for the variable Rate, which means when Rate, IMDb Rating, increases by one point while holding all other variables constant, we expect the number of votes for a film to increase by 104941.1. With Duration we receive a slope of 1817.8, which means that for every one minute increase in Duration, holding all other variables constant, we expect the number of votes for a film to increase by 1817.8. With the variable Violence(Mild) we obtain a value of 48577.2, which means that using a baseline/reference of Violence being None we expect the number of votes for a film to increase by 48577.2 when the Violence level is instead Mild. A Mild level of Violence number of votes for a film is greater than the None level of violence number of votes for a film by 48577.2. With the variable Violence(Moderate) we obtain a value of 93501.2, which means that using a baseline/reference of Violence being None we expect the number of votes for a film to increase by 93501.2 when the Violence level is instead Moderate. A Moderate level of Violence number of votes for a film is greater than the None level of Violence number of votes for a film by 93501.2. With the variable Violence(Severe) we obtain a value of 85563.9, which means that using a baseline/reference of Violence being None we expect the number of votes for a film to increase by 85563.9 when the Violence level is instead Severe. A Severe level of Violence number of votes for a film is greater than the None level of Violence number of votes for a film by 85563.9.

We observe that for each level of Violence the number of votes for a film increases using None as a baseline/reference, meaning films with a None level of Violence receive the least number of votes in comparison to other levels.

We also observe that for each variable in our model, they all have extremely small p-values which are less than the significance level of 0.05. Rate and Duration have p-values < 2e-16, Violence(Mild) has a p-value of 0.000115, Violence(Moderate) has a p-value of 9.12e-14, and Violence(Severe) has a p-value of 5.14e-11, all less than the significance level of 0.05. Therefore, all of the predictors are statistically significant and are significant predictors at the 5% significance level. The model itself has a p-value of 2.2e-16, under the significance level of 0.05. Because the p-value is less than $\alpha = 0.05$, we conclude that the independent variables do significantly predict Votes.

From the coefficient of determination, our predictors Violence, Duration, and Rate are able to account for approximately 28.99% of the observed variance in number of votes. We used the adjusted r-squared value since this is a multiple linear regression model.

We can now proceed to testing the final assumptions.
Final Assumptions:
Linearity

```
raintest(final)
```

```
##
##  Rainbow test
##
## data:  final
## Rain = 3.2117, df1 = 1613, df2 = 1606, p-value < 2.2e-16
```

Using the raintest to test the linearity assumption we receive a p-value of nearly 0, specifically 2.2e-16, and the linearity assumption fails.

Normality of residuals

```
resid <- residuals(final)
shapiro.test(resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.82636, p-value < 2.2e-16
```

Observing the shapiro test for the residuals, we receive a p-value of nearly 0, specifically 2.2e-16. We can conclude that the residuals are not normally distributed and the normality of residuals assumption fails.

Homoscedasticity & Linearity

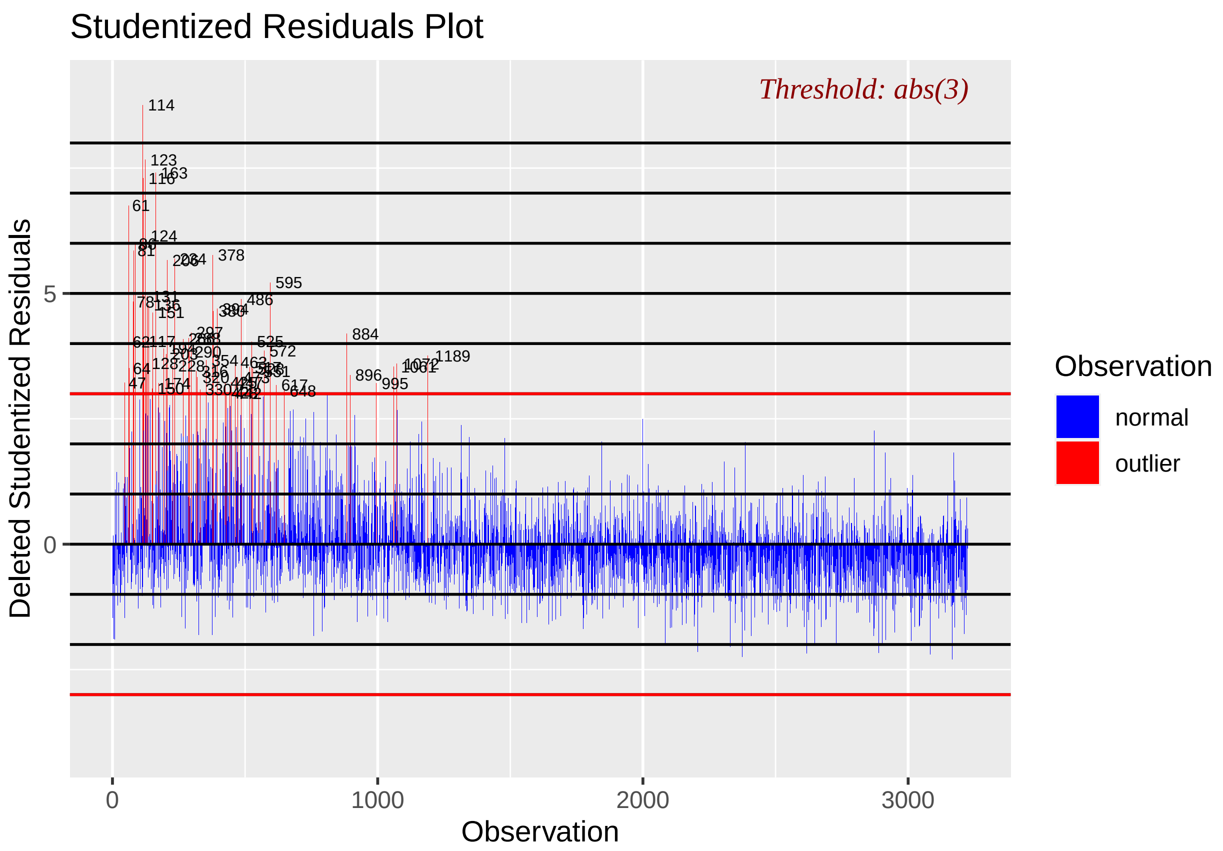
```
ols_test_breusch_pagan(final)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  ---------------------------------
##  Response : Votes
##  Variables: fitted values of Votes
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    2476.5480
##  Prob > Chi2   =    0.0000
```

For the homoscedasticity (equal variance of residuals) assumption we obtained a p-value of nearly 0, and the assumption does not hold.

Outliers

```
ols_plot_resid_stud(final)
```



We observe that there are outliers within our model towards the left side of the plot. There are multiple significant outliers outside of the -3 to 3 threshold, and the outlier assumption is not met.

Multicollinearity

```
vif(final)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## Rate                    1.192808  1        1.092157
## Duration                1.212119  1        1.100963
## as.factor(Violence) 1.038888  3        1.006379
```

The multicollinearity assumption is met, as all values obtained in the model are under the threshold of 5 and no variables are too highly correlated with each other.

Autocorrelation

```
dwtest(final)
```

```
##
##  Durbin-Watson test
##
## data:  final
## DW = 1.6852, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

We observe a p-value of nearly 0, 2.2e-16, so we can reject the null hypothesis and the residuals are not autocorrelated.
We observe that many of the final assumptions are not met, except for multicollinearity. We can conclude that this model is not a very strong predictor for the number of Votes and shouldn't necessarily be used. We unfortunately conclude that this model is not very useful in real world situations since many important assumptions are violated. We did assume normality for our dependent variable, which may have been a factor in many of these assumptions failing.

Section 4: Conclusion and Reccommendation

In our analysis, we began by observing at the relationship between Profanity and Violence in relation to total number of votes to see if this could help us when modeling later and to observe if there was any association between the variables. Before we began analysis, we observed that movies with higher levels of Profanity tended to have higher levels of Violence within our plot. Conducting our analysis we observed that the two variables were associated. This told us that the different categories of movie content could be correlated, so it could help us make a simpler model. This is consistent with the literature, which finds that violence and profanity are typically correlated. Furthermore, violence correlates with higher performance at the box office, which means a movie is more popular (Switzer and Lang, 2008). We also ran a post-hoc test analyzing the Violence level of 'None' to the total proportions of each level of Profanity. We observed a statistically significant difference in proportions between three of the six groups, those groups being None-Moderate, None-Severe, and Mild-Severe proportions. It is interesting to see that the levels None were statistically significant with respect to the two highest possible levels, moderate and severe. Using our results and model, we observe that the proportion of number of votes between the levels of Violence and Profanity increases as the levels increase together. We suggest that if you want to achieve a high level of either category in your film, you have have a high level of the other variable Profanity or Violence as they are associated with one another. We also suggest possibly exploring a more in depth association between the variables, since the variables in this data set have 4 generic levels which are are not a completely accurate measurement of Violence and Profanity in films.

Next, we looked into how the Frightening variable is related to log Votes. There was a statistically significant difference in the numbers of log Votes between the levels of Frightening. We concluded that audiences are more drawn to certain types of content in movies, which supports the idea that other "content" variables such as Violence or Profanity may also have a significant effect on the number of votes. We observed with the Tukey's test which levels were statistically significant with each other, those levels being Moderate-None, Severe-None, and Moderate-Mild. Again, relating to the previous question, we see an association with the two greatest levels of Frightening, Severe and Moderate, with the lowest level, None, with the higher level of Frightening having a greater mean number of log votes both times. This means that a greater level of Frightening corresponded to a greater number of log votes, in comparison to movies that didn't seem very frightening, something to consider when making films. We suggest with these results using a greater level of Frightening in order to achieve a greater number of votes. This can't necessarily apply to everything however, as one important limitation to this idea is that children movies most likely shouldn't be very frightening.

Finally, we created a linear model which able to predict the amount of votes to some extent, but the most popular movies are unpredictable. We see this effect in our plots of the variables and the residuals. Our model unfortunately failed many final assumptions, meaning it isn't necessarily that good at predicting number of votes in reality. There seems to be a threshold for the number of votes where past that point it is difficult to predict what factors are influencing the popularity of a movie. Our predictors were only able to significantly account for approximately 28.99% of the variation in number of Votes, so there is a majority that is unaccounted for. There are likely other factors outside of this data set that could warrant further investigation, such as strong negative or positive reviews, reviews from critics, as well the presence of recognized stars in the cast (Carrillat et. al., 2018). We suggest that model should not be used in the real world to predict number of votes a film receives due to the failure of many assumptions and model coefficients that aren't realistic to the real world.

One major limitation we had was the variables within our data set. We suggest further research by exploring the factors mentioned above. There are many other factors outside of the variables we had that can influence the number of votes a movies receives and the popularity of the movie. Furthermore, our data is restricted to ratings found on IMDb, so it is difficult to generalize our finding to other major rating websites, such as Google or Rotten Tomatoes, which also have very large user bases and are thus important to the movie industry when getting information about what customers think about various movies.

References

Abdi and Williams, 2010, https://personal.utdallas.edu/~Herve/abdi-NewmanKeuls2010-pretty.pdf

Allotey, 2022, Multiple-Linear-Regression.pdf

Carrillat et. al., 2018, https://link.springer.com/article/10.1007/s11747-017-0561-6

Data (Kaggle): https://www.kaggle.com/datasets/mazenramadan/imdb-most-popular-films-and-series

Lee et. al., 2014, https://ieeexplore.ieee.org/document/6741434

Pentheny, 2015, https://scholars.unh.edu/cgi/viewcontent.cgi?article=1267&context=honors

sthda, 2018, http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

Switzer and Lang, 2008, https://repository.stcloudstate.edu/cgi/viewcontent.cgi?article=1003&context=econ_seminars)