Predicting and Exploring IMDB's Ratings

Juan Borgnino (jb3852)
Carolyn Morris (cm3491)
Jose Ramirez (jdr2162)
Manuel Rueda (mr3523)

December 15, 2015

Abstract

In this analysis, we take IMDB's movie rating and descriptive data, combine it with the Academy Awards Best Picture nominations, and derive a model to predict highly rated movies. For this we test linear, polynomial, splines and General Additive Models. We identify our dataset as a complex one to model, so more flexible models performed better. In terms of inference, we found the number of votes was the best predictor, implying that more popular movies tend to be rated higher (or the other way around). Other variables such as the year of release, its length, budget and genre are also significantly related. Contrary to our expectations, 'Best Picture' nominations were not.

Data Sources

Our dataset consists of movie rating and budget data for 5,183 films from the Internet Movie Database, paired with a historical list of Academy Awards Best Picture Nominations, retrieved from Agg Data. Both data sources are legally made available to the general public. For our analysis, we are interested in analyzing the interactions between IMDB ratings assigned by the general public and a list of other possible explanatory variables, which are listed on the following table.

Variable	Description
title	Title of the movie.
year	Year the movie was released.
budget	Total budget if in US dollars.
length	Length of movie (in minutes).
rating	Average IMDb user rating.
votes	Number of IMDb users who rated the movie.
mpaa	MPAA rating.
nominated	Binary variable indicating if the movie was nominated for the 'Best Movie'.
genre	Binary variables indicating whether movie belongs to any of the following genres: action, animation, comedy, drama, documentary, romance, short.

Exploratory Data Analysis & Tests

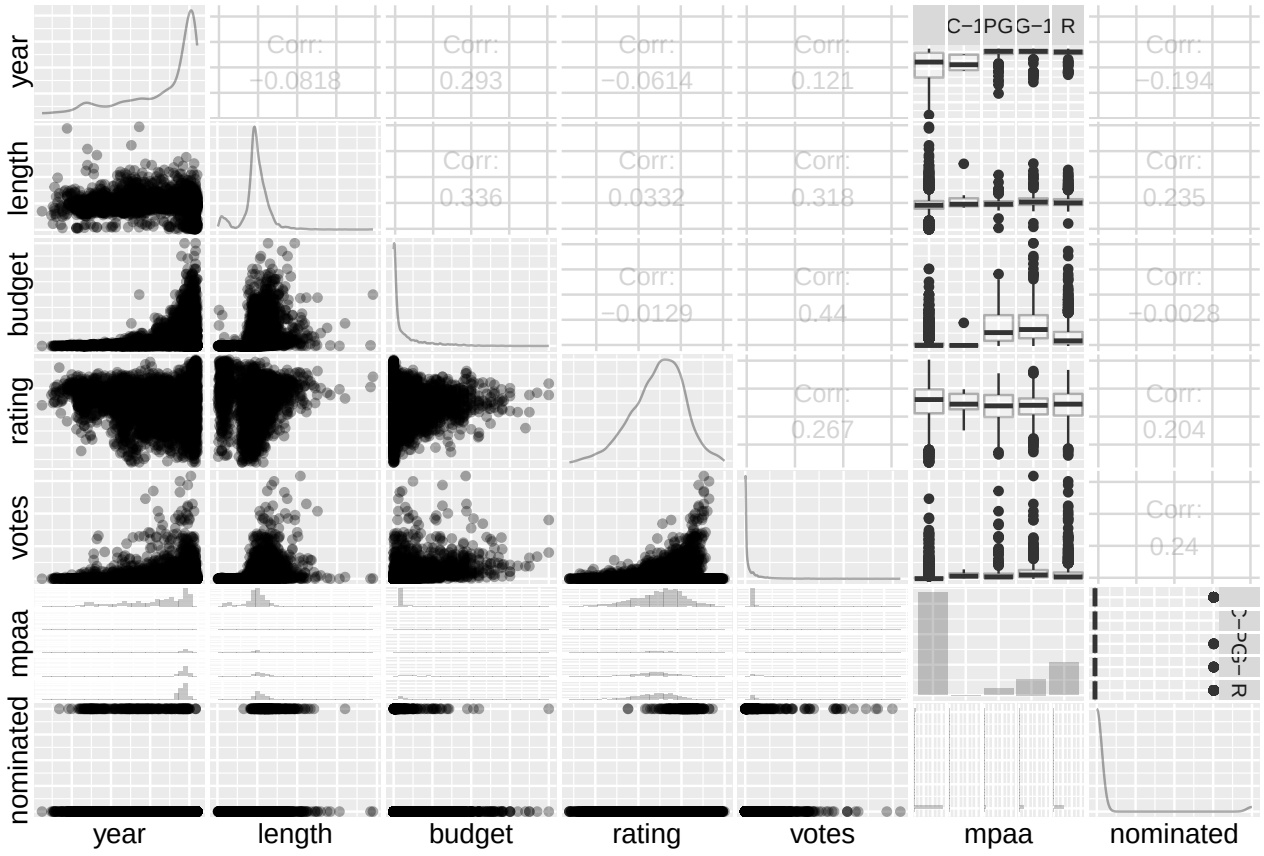
Summary of Data

Before beginning with the formal statistical analysis, it is convenient to perform exploratory analysis on the data to identify possible patterns. We begin by looking at the summary information for each of the variables.

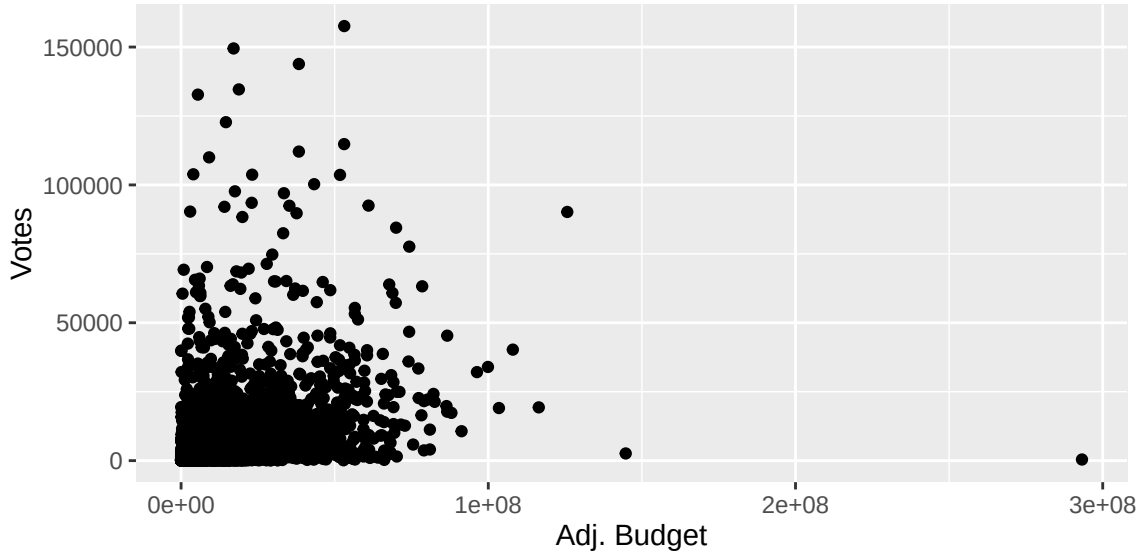
```
##     title                year          length            budget
##  Length:5183        Min.   :1906   Min.   :  1.00   Min.   :1.00e+03
##  Class :character   1st Qu.:1975   1st Qu.: 86.00   1st Qu.:2.75e+05
##  Mode  :character   Median :1996   Median : 97.00   Median :3.00e+06
##                     Mean   :1985   Mean   : 96.41   Mean   :1.35e+07
##                     3rd Qu.:2001   3rd Qu.:112.00   3rd Qu.:1.50e+07
##                     Max.   :2005   Max.   :390.00   Max.   :2.00e+08
##      rating           votes           mpaa          Action
##  Min.   : 1.000   Min.   :     5        :3370   Min.   :0.0000
##  1st Qu.: 5.200   1st Qu.:    70   NC-17:   7   1st Qu.:0.0000
##  Median : 6.300   Median :   622   PG   : 212   Median :0.0000
##  Mean   : 6.137   Mean   :  5004   PG-13: 530   Mean   :0.1609
##  3rd Qu.: 7.200   3rd Qu.:  4682   R    :1064   3rd Qu.:0.0000
##  Max.   :10.000   Max.   :157608                Max.   :1.0000
##    Animation           Comedy           Drama         Documentary
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.02392   Mean   :0.3373   Mean   :0.4553   Mean   :0.02392
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##     Romance           Short          nominated
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.1488   Mean   :0.08065   Mean   :0.04245
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
```

We see a couple of interesting things here:

(1) 25% of the movies have received less than 70 *votes*. We will remove this lower quantile, as we want to focus our attention on those films for which a larger consensus has been reached.
(2) In the *length* column (below), we can see a cluster of short films on the right hand side of the plots. They clearly have a different behavior than full-feature ones, so we will exclude them from the analysis.
(3) The *mpaa* variable has 4 categorical variables, each corresponding to a rating (e.g. PG-13). We will replace this with *mpnum* (0:N/A, 1:PG, 2:PG-13, 3:NC-17, 4:R).
(4) The *budget* variable is defined on US dollars, but has not been adjusted for inflation. We will retrieve the yearly Consumer Price Index (CPI) data to convert this series into real terms.

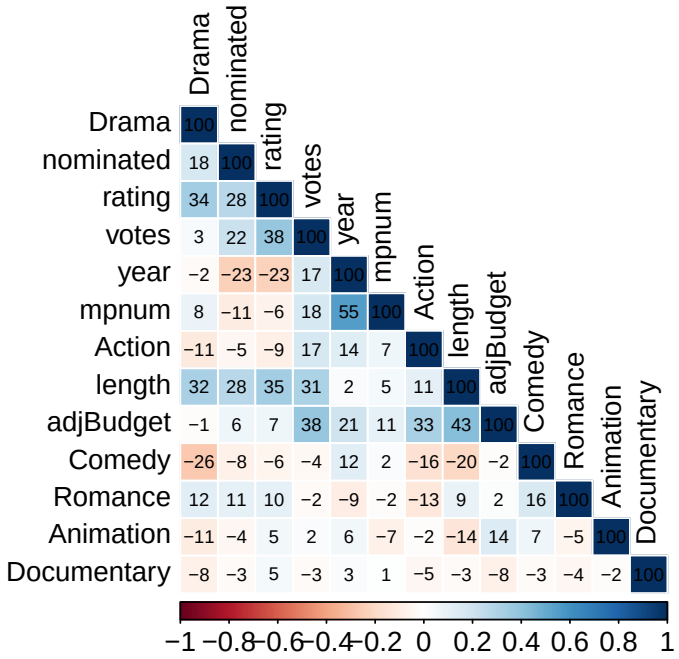


After proceeding with these changes, we identified one extreme outlier in terms of real budget: "Voyna i mir", or "War & Peace" (1966). Investigating this movie, we see that it was sponsored by the Soviet party, which covered its astronomical expenses. For the purposes of the analysis, we will omit it from the sample.

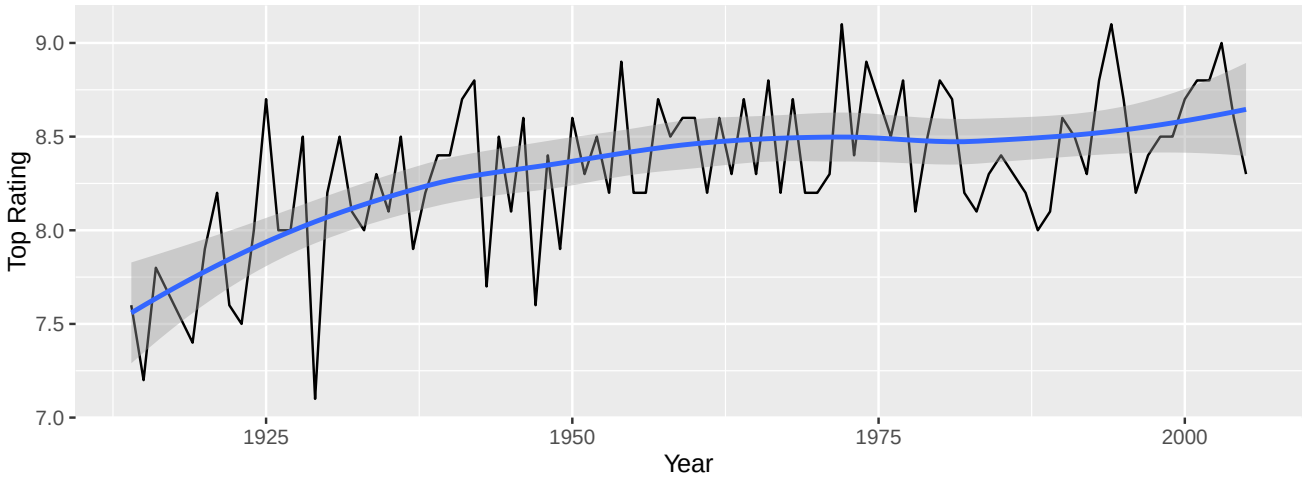


Correlations Across Variables

Now we present the charts of paired variables, along with a visual representation of the correlation matrix.



Here we can see a positive correlation between rating and *votes*, as well as rating and *nominated*. The correlation with *adjBudget* is surprisingly low, suggesting that this is not one of the main driving factors on explaining the user preference for a film. *Year* is also negatively correlated, although on the pairs plot this seems somewhat contradictory; the older movies have a rating concentration on the upper part of the spectrum. This is explained by the high density of recent movies that received rating: if we take the upper bound of ratings per year, we see that it increases with time:



Finally we see *length* positively correlated with *rating*, so we can expect longer movies to receive more love from the audience.

Other interesting observations:

- Action & Animation movies tend to receive a high budget, which makes sense considering they rely intensively on special effects.

- Dramas and Romance films tend to be nominated more frequently.
- Dramas and Comedies don't tend to go together, while Comedies and Romance are a frequent mix (Rom-Coms).

We are however mostly interested in how the IMDB rating of a movie can be predicted utilizing the rest of the variables, so we will restrict our attention to these specific interactions. An interesting exercise is to investigate how the ratings correlate with the rest of the variables, particularly the most significant ones (*votes, adjBudget, & nominations*) when splitting the dataset by genre, and for this we will make use of the **Test of Independence**, which is defined as follows:

Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. We wish to test the hypothesis that X and Y are independent. We will use R, the sample estimate of ρ, to test the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$. For this we build the following statistic, which has a t-distribution with $n - 2$ degrees of freedom.

$$T = \left(\frac{R\sqrt{n-1}}{\sqrt{1-R^2}}\right)$$

We reject the null hypothesis with a level α if $|T| \geq t_{\alpha/2,n-2}$. This is equivalent to applying the `cor.test()` R function, however for completeness we have built our own function that will lead us to the same results.[1] In our case we don't have evidence of normality in our variables, but we will assume so for practical purposes. Below are the results, per genre.

```
## [1] "rating vs. votes"
```

```
##         genre         r         t    t-crit    n  sig?      p-value
## 1      Action 0.4946832 15.241589 -1.963278  719  TRUE 0.000000e+00
## 2   Animation 0.6246729  7.959437 -1.984217  101  TRUE 1.465716e-12
## 3      Comedy 0.4074920 16.588342 -1.961682 1384  TRUE 0.000000e+00
## 4       Drama 0.4017999 18.959011 -1.961235 1869  TRUE 0.000000e+00
## 5 Documentary 0.2876246  1.875466 -2.022691   41 FALSE 3.411472e-02
## 6     Romance 0.2892363  7.867535 -1.963469  680  TRUE 7.105427e-15
```

```
## [1] "rating vs. budget"
```

```
##         genre            r          t    t-crit    n  sig?      p-value
## 1      Action  0.161332717  4.3773255 -1.963278  719  TRUE 6.901112e-06
## 2   Animation  0.208695064  2.1232418 -1.984217  101  TRUE 1.811337e-02
## 3      Comedy -0.027537510 -1.0241025 -1.961682 1384 FALSE 8.470170e-01
## 4       Drama  0.005046193  0.2180427 -1.961235 1869 FALSE 4.137098e-01
## 5 Documentary -0.246881644 -1.5910245 -2.022691   41 FALSE 9.401608e-01
## 6     Romance -0.042133538 -1.0980664 -1.963469  680 FALSE 8.637174e-01
```

```
## [1] "rating vs. oscar"
```

```
##         genre         r         t    t-crit    n sig?      p-value
## 1      Action 0.2373732  6.543120 -1.963278  719 TRUE 5.732681e-11
## 2   Animation        NA        NA -1.984217  101   NA           NA
## 3      Comedy 0.2133804  8.119471 -1.961682 1384 TRUE 4.440892e-16
## 4       Drama 0.3319948 15.207656 -1.961235 1869 TRUE 0.000000e+00
## 5 Documentary        NA        NA -2.022691   41   NA           NA
## 6     Romance 0.3424068  9.489352 -1.963469  680 TRUE 0.000000e+00
```

[1] Check code appendix for details.

The main observations from this analysis:

- The number of votes are highly correlated with the movie rating in all cases, except documentaries, where we have few observations.
- Action and Animation movies tend to perform better in terms of rating when they have a higher budget. This is in line with our previous observations.
- Movies that have been nominated for Best Picture also tend to receive better ratings. `NA`'s appear because some genres have not received this nomination at all.

Distribution of Ratings

We have seen that ratings across different genres exhibit different correlations, and now we wish to investigate with more detail the distribution of votes among these different categories. First, we will comment on the median of this variable; can we say that it is below the centered-rating of 5 stars? A box plot of the data and the **Sign-Test** can help us investigate this.

For the Sign-Test, we will test the following hypothesis:

$$H_0 : \theta < 5 \text{ vs. } H_1 : \theta \geq 5$$

The test can be performed via the `SIGN.test` function in R, but here we have decided to implement the function step by step.[2]

```
## [1] "Ratings subset:"
```

```
##  [1] 4.3 4.5 4.7 4.9 5.1 5.4 5.4 5.5 5.9 5.9 6.2 6.2 6.3 6.5 6.5 6.6 6.6
## [18] 6.6 6.7 6.7 6.9 7.0 7.6 7.7 7.7 7.9 8.0 8.2 8.7 8.7
```

```
## [1] "Length of subset:"
```

```
## [1] 30
```

```
## [1] "Observations greater than 5:"
```

```
## [1] 26
```

```
## [1] "p-value:"
```

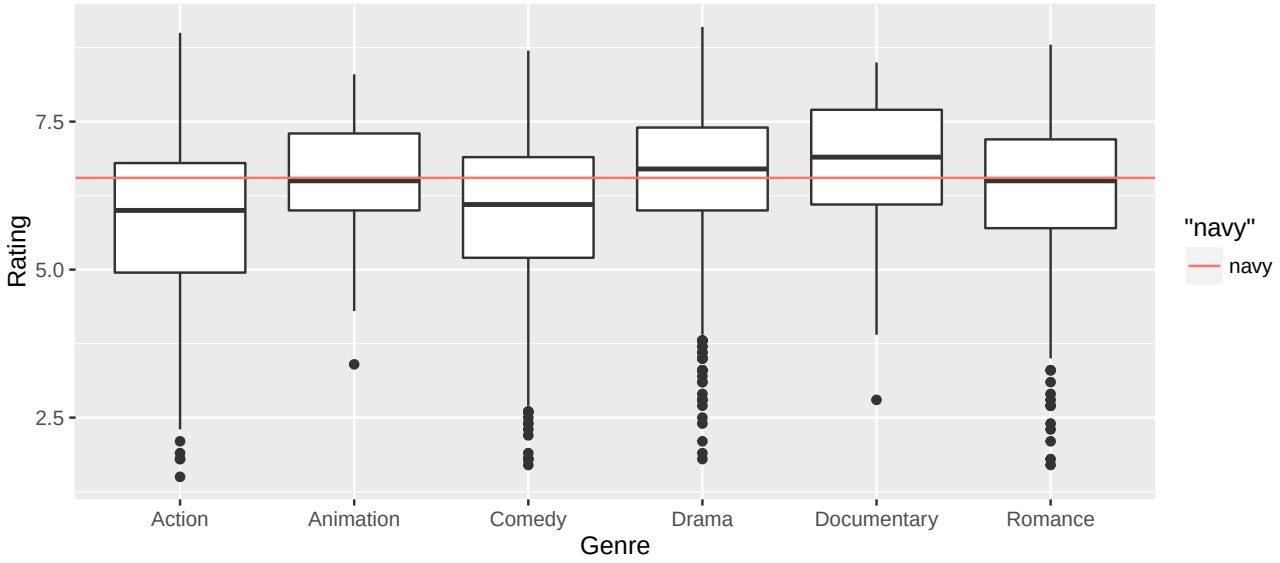
```
## [1] 2.973806e-05
```

```
## [1] "sample median:"
```

```
## [1] 6.55
```

[2] Check code appendix for details.

With this small *p-value*, we reject the null hypothesis of the median being less than 5 stars. Combining this information with a box plot:



The graphical representation also confirms what the sign-test suggested. We see that votes tend to be concentrated on the upper part of the spectrum, with a median around 6.5. Most movies tend to be "good" ones, or they tend to be rated more. Now, we see some difference on the distributions across genres. We now wish to test if this difference is statistically significant. For this we will use the **Mann–Whitney–Wilcoxon** statistic on a couple of pairs, which will test for:

$$P(Y \leq y) = P(X + \Delta \leq y) = F(y - \Delta)$$

$$H_0 : \Delta = 0 \text{ vs. } H_1 : \Delta \geq 0$$

Under H_0, the distributions of X and Y are the same, and we can combine the samples to have one large sample of $n = n1 + n2$ observations. We will use the `wilcox.test()` command in R to perform this test. Below: statistic & P-Value.

```
## [1] "Drama & Comedy"
```

```
##       [,1]         [,2]
## W 938324.5 5.298633e-41
```

```
## [1] "Comedy & Romance"
```

```
##       [,1]         [,2]
## W 380133.5 1.177146e-12
```

```
## [1] "Action & Comedy"
```

```
##     [,1]       [,2]
## W 465176 0.01422705
```

First, looking at Drama and Comedy, we see that indeed their distributions are statistically different, which was also suggested by the box plots, and we can relate back to the negative correlation observed between the variables (i.e. few Dramatic Comedies exist). Romance and Comedy however do tend to go together, but still the WHM

test shows us that their distributions are different. Even for Action and Comedy, two genres on which the box plot suggest a similar distribution, we do not find the null hypothesis significant at $\alpha = 0.01$.

Given this, it is convenient to keep a distinction between movies of different categories. There are two ways to achieve this:

- Subset the data into different groups, one per genre.
- Keep binary variables to identify when a given film belongs to a specific genre.

We have decided to take the second approach, since this way we don't reduce the sample size of categories such as Documentaries, and we can also account for movies on which more than one genre is present (i.e. Action + Animation, Romance + Comedy, etc.).

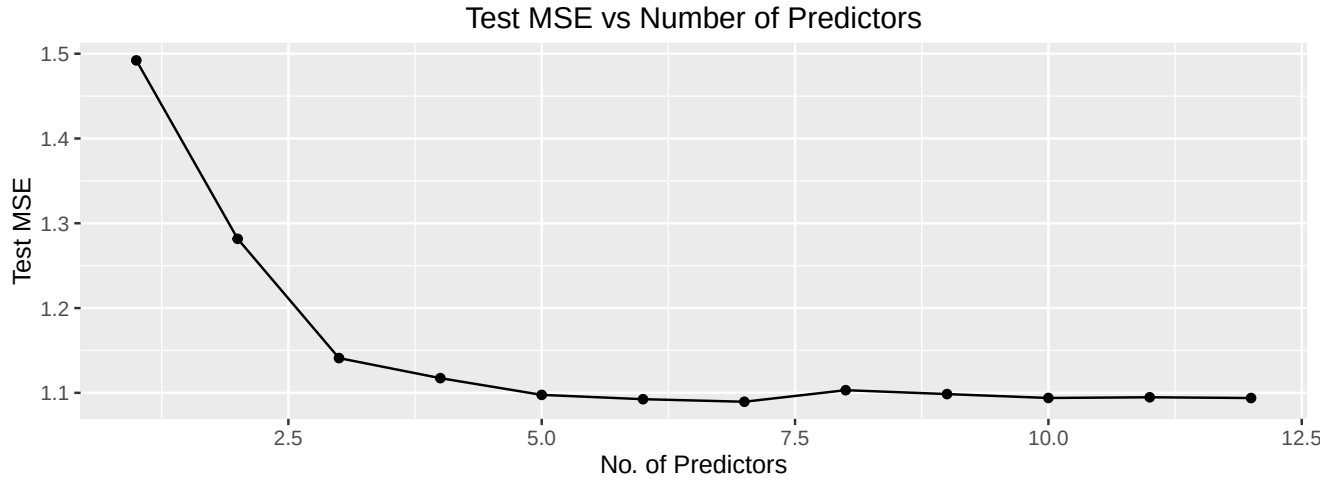
Statistical Modeling of Ratings

Having completed the exploratory analysis of the data, we proceed to apply a number of statistical modeling techniques to predict the response rating. We first divide our data set into a training set, using 80% of the data, and test set. We will start by applying best subset selection, and then we will fit several models. For each model, we will provide a brief explanation of the method and a summary of the results obtained.

Best Subset Selection

Given that we have 12 predictors, we use **Best Subset Selection** to decide what variables to include. This method allows us to fit a separate least squares regression for each possible combination of the 12 predictors. Hence, we will fit in total $2^{12} = 4096$ models and choose the one with smallest Test MSE. It is important to notice that if the number of predictors was greater, we would have used **Forward Subset Selection** due to the lower computational costs.

The following plot shows the test MSE for each of the separate best models of all sizes up to 12 in terms of RSS.



Linear Regression

Next, we present a summary of the regression using the model obtained using **Best Subset Selection**. The included variables are: year, length, votes, Animation, Comedy, Drama, and adjBudget. Somewhat surprisingly, *nominated* was not found to be relevant, which could be explained by the fact that only a small fraction of the movies (5.66%) actually received a nomination.

From the summary we observe the following points:

- All p-values are smaller than 0.01 . Hence, we can reject the null hypothesis and conclude that the predictors are statistically significant and are related to the response.
- We obtain an adjusted R squared value of 0.394, meaning that our model is explaining aprox. 40% of the variance of the rating.
- The Test MSE is equal to 1.089. That is, on average our prediction is missing the true value of the response by 1.089.

Some of the interpretation of the coefficients are:

- A 1000 increase in votes, increases, on average and given that all other predictors are fixed, the rating by 0.041.
- A 1,000,000 dollar increase in the budget of the movie, reduces, on average and given that all other predictors are fixed, the rating by -0.014.

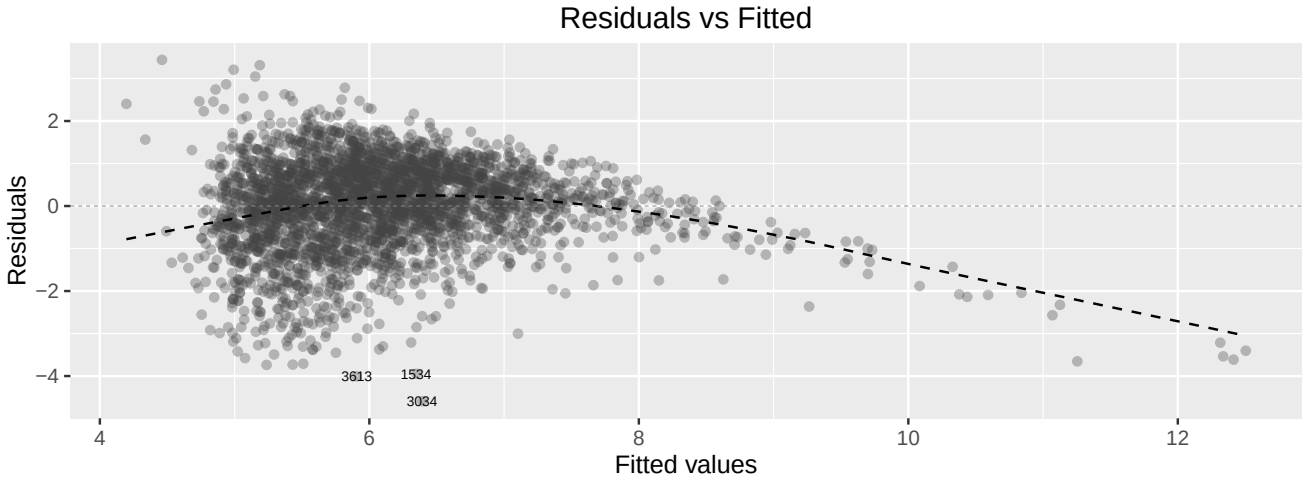
```
## [1] "Test MSE:"
```

```
## [1] 1.089497
```

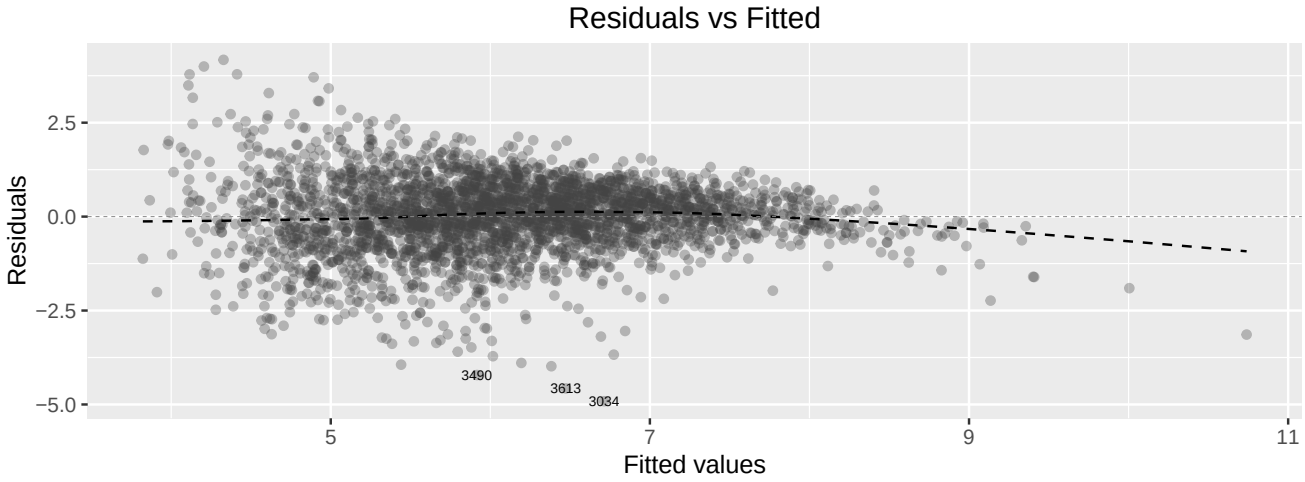
```
##
## Call:
## lm(formula = rating ~ year + length + votes + Animation + Comedy +
##     Drama + adjBudget, data = train.set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5928 -0.5748  0.1089  0.7162  3.4389
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.148e+01  1.918e+00  21.632  < 2e-16 ***
## year        -1.888e-02  9.643e-04 -19.582  < 2e-16 ***
## length       1.408e-02  9.612e-04  14.652  < 2e-16 ***
## votes        4.127e-05  1.606e-06  25.696  < 2e-16 ***
## Animation    1.218e+00  1.261e-01   9.657  < 2e-16 ***
## Comedy       3.236e-01  4.098e-02   7.897 3.94e-15 ***
## Drama        7.359e-01  4.094e-02  17.978  < 2e-16 ***
## adjBudget   -1.353e-08  1.453e-09  -9.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 3087 degrees of freedom
## Multiple R-squared:  0.3961, Adjusted R-squared:  0.3948
## F-statistic: 289.3 on 7 and 3087 DF,  p-value: < 2.2e-16
```

We now look at several plots of the residuals in order to detect potential problems from our linear model.

From the graph we find that the residuals exhibit a clear inverted U-shape, which provides a strong indication of non-linearity in the data. Therefore we should use some non-linear transformation of the predictors.



In order to take account for the non-linearity we found, we try taking the logarithm of the variable votes. Below, we plot the new residuals and confirm there is now no discernable pattern in the distribution of the errors. Hence, the true model seems to be non-linear.



Finally, we substitute from our model the variable votes for the logarithm of votes and check if our models improves. From the new fit we find the following positives results:

- The Test MSE has slightly decreased to 1.037.
- The Adjusted R Squared increased to 0.478.

To conclude, using the logarithm of votes seems to be improving the fit of our model. Hence, for the rest of the statistical learning methods we will apply, we will use this variable.

```
## [1] "Adj. R Squared:"
```

```
## [1] 0.4786333
```

```
## [1] "Test MSE:"
```

```
## [1] 1.03734
```

Polynomial Regression

We extend the linear regression model by replacing it with a polynomial function. That is, some of our predictors we will raised to a power. We determine which degree polynomial to implement with the Analysis of Variance (ANOVA).

ANOVA performs a hypothesis tests using an F-test to compare different models. The null hypothesis is that the two models are equally good explaining the data; the alternative hypothesis is that the more complex model has higher explanatory power. It is important to consider when performing the test that the models must be nested, i.e. model M1 is a subset of M2. If the p-value is greater than 0.05, then there is not enough evidence to reject the null, and we conclude that the simpler model, M1, is sufficient to explain the data.

Next we use ANOVA to test on the model we obtained from the **Linear Regression** section the degree of polynomial to fit our model. Below we observe the results from our test.

- With an F-statistic of 16 and p-value very close to zero, we have strong evidence to believe that the model which includes the cubic of the logarithm of votes is better than the model that only includes the logarithm of votes.

```
## Analysis of Variance Table
##
## Model 1: rating ~ year + length + Animation + Comedy + Drama + adjBudget +
##     log.votes
## Model 2: rating ~ year + length + Animation + Comedy + Drama + adjBudget +
##     poly(log.votes, 2)
## Model 3: rating ~ year + length + Animation + Comedy + Drama + adjBudget +
##     poly(log.votes, 3)
## Model 4: rating ~ year + length + Animation + Comedy + Drama + adjBudget +
##     poly(log.votes, 4)
## Model 5: rating ~ year + length + Animation + Comedy + Drama + adjBudget +
##     poly(log.votes, 5)
##   Res.Df    RSS Df Sum of Sq        F    Pr(>F)
## 1   3861 3684.0
## 2   3860 3578.0  1   106.000 115.0519 < 2.2e-16 ***
## 3   3859 3563.2  1    14.807  16.0709 6.217e-05 ***
## 4   3858 3557.7  1     5.484   5.9524   0.01474 *
## 5   3857 3553.6  1     4.150   4.5045   0.03387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, to confirm that the polnomial of degree 3 is the best model so far, we predict on our test set. We obtain a Test MSE of 0.991, which is an important improvement from the model without the cubic polynomial.

```
## [1] "Test MSE:"
```

```
## [1] 0.9915019
```

Before moving on to apply other methods, we take a closer look at the variable *votes*, since it seems to be a very important predictor for the response. We regress rating onto the cubic of the logarithms of votes and observe the following results:

- The adjusted R squared is aprox 19%. Hence, this variable alone seems to be doing a good job explaining almost 20% of the variance of rating.
- We obtain a test MSE of 1.4607. This a very good result considering that using all 7 variables we obtained a result of 0.991.

```
## [1] "Adj. R Squared:"
```

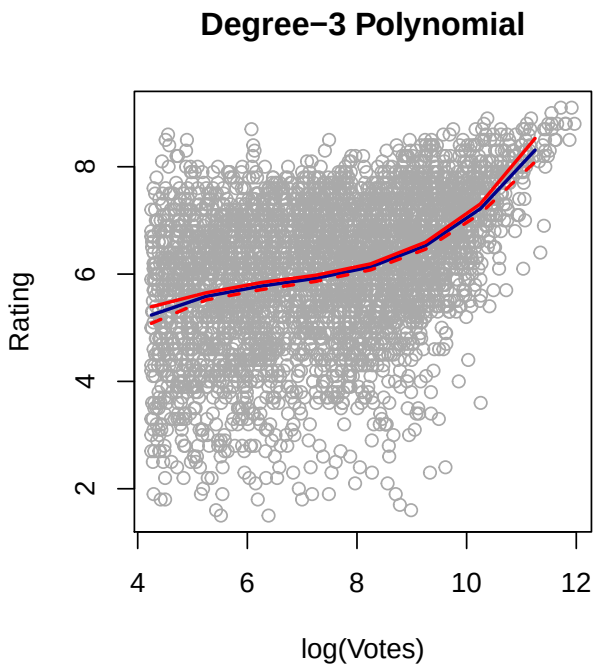
```
## [1] 0.1866167
```

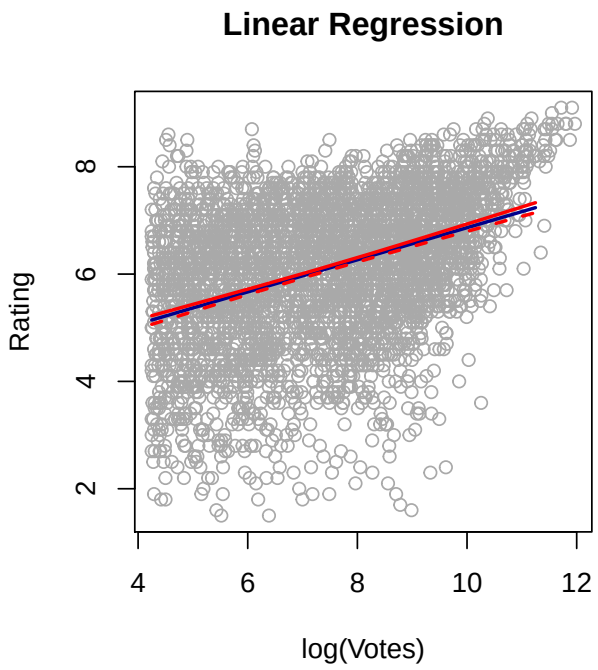
```
## [1] "Test MSE:"
```

```
## [1] 1.46075
```

To compare how linear regression and the cubic polynomial models only using the logarithm of votes fit the data we plot both curves. The first graph is produced by fitting a polynomial of degree 3, while the graph on the right is produced by fitting a regular line. In both graphs, the blue line represents the fitted value from both models and the red lines an estimated 95% confidence interval. It is very clear how the polynomial of degree 3 allows us to fit a more flexible curve that fits the data better.

```
##
## Call:
## lm(formula = rating ~ poly(log.votes, 3), data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8009 -0.7590  0.1270  0.8831  3.2342
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.10408    0.01935 315.445  < 2e-16 ***
## poly(log.votes, 3)1 33.31060    1.20364  27.675  < 2e-16 ***
## poly(log.votes, 3)2 11.21621    1.20364   9.319  < 2e-16 ***
## poly(log.votes, 3)3  7.39125    1.20364   6.141 9.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.204 on 3865 degrees of freedom
## Multiple R-squared:  0.1872, Adjusted R-squared:  0.1866
## F-statistic: 296.8 on 3 and 3865 DF,  p-value: < 2.2e-16
```





Regression Splines

Regression splines allow us to fit separate low-degree polynomials over different regions of our predictor. It is important to note the difference from the polynomial regression we just applied. Before, we imposed a global cubic polynomial to the data and now we will implement a piecewise polynomial to k knots (the points were the coefficients change). Therefore, it is clear that regression splines allow us to fit more flexible and complex models to our data.

Cubic Spline

We start by fitting a cubic spline to our data. We keep on using the logarithm of votes as our predictors since we obtained the best results using this variable. Also, keeping the same predictor, we can compare all the models.

In order to determine the degrees of freedom, we use cross-validation to obtain an estimate of the test error for different degrees of freedom. We then choose the number which minimizes the error. We find that 7 degrees of freedom provide us the lowest estimated test error. This means that we will use 4 knots.

The 4 points were we will fit the knots are:

```
##      20%      40%      60%      80%
## 5.615314 6.953875 8.092665 9.151567
```

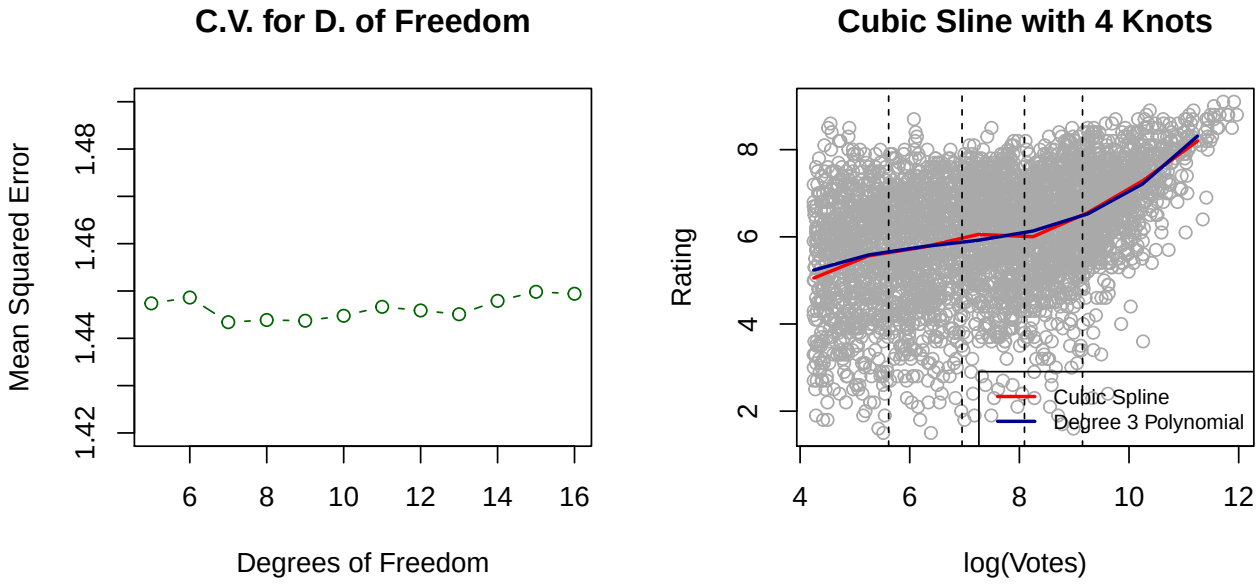
Next we fit a spline to the data. From the summary of the model we find the following results:

- Almost all coefficients are statistically significant and related to the response.
- The Adjusted R Squared is 0.1905: slightly higher than the cubic polynomial.

```
## [1] "Adj. R Squared:"
```

```
## [1] 0.190519
```

Below we can observe a graph showing the estimated Test MSE for each degree of freedom. The graph on right shows how cubic spline has some additional flexibility compared the cubic polynomial.



In last place, we obtain a test MSE of 1.464 from the cubic spline model, that is slightly higher than the polynomial regression. It looks likes the additional flexibility could be increasing the error due to higher variance.

```
## [1] "Test MSE:"
```

```
## [1] 1.464965
```

Natural Cubic Spline

A natural spline is a regression spline with an additional constraint: the function is required to be linear in the boundaries. The advantage of using a natural cubic spline is that usually in the boundaries there are less points. Hence, a polynomial can be quite unstable in the boundaries.

Like before, we start by using cross-validation to obtain the degrees of freedom which minimize the test error estimate.

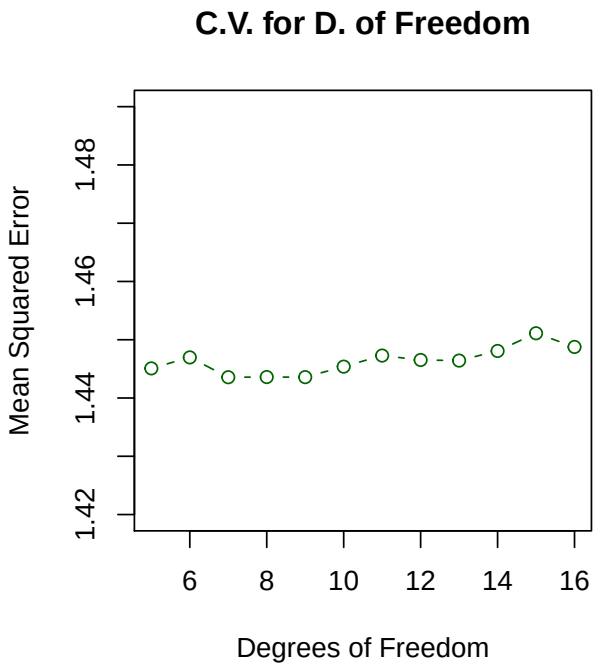
We fit a natural cubic splin with 6 knots since we obtained 7 degrees of freedom from cross validation. From the summary of the fit we observe:

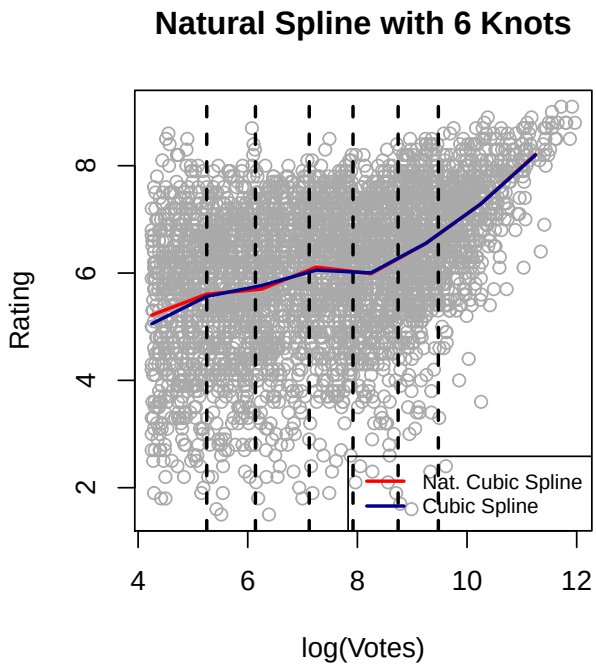
- Almost the same adjusted R squared, 0.1908 than what we obtained from the cubic spline.
- Almost all coefficients are statistically significant.

```
##
## Call:
## lm(formula = rating ~ ns(log.votes, df = ns.best.df), data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7906 -0.7578  0.1265  0.8754  3.2260
##
## Coefficients:
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       5.2155     0.1047  49.828  < 2e-16 ***
## ns(log.votes, df = ns.best.df)1   0.2508     0.1285   1.952 0.051046 .
## ns(log.votes, df = ns.best.df)2   1.1491     0.1660   6.924 5.12e-12 ***
## ns(log.votes, df = ns.best.df)3   0.5368     0.1411   3.806 0.000144 ***
## ns(log.votes, df = ns.best.df)4   1.0179     0.1391   7.319 3.02e-13 ***
## ns(log.votes, df = ns.best.df)5   1.6034     0.1418  11.306  < 2e-16 ***
## ns(log.votes, df = ns.best.df)6   3.3188     0.2809  11.813  < 2e-16 ***
## ns(log.votes, df = ns.best.df)7   3.4606     0.2435  14.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.201 on 3861 degrees of freedom
## Multiple R-squared:  0.1922, Adjusted R-squared:  0.1908
## F-statistic: 131.3 on 7 and 3861 DF,  p-value: < 2.2e-16
```

In addition, we graph both models and observe how the natural cubic spline has a little more flexibility due to the higher number of knots. However, we do not observe any improved performance. Particularly because in the boundaries there is still a large number of observations. As a result, the cubic spline is as stable as the natural cubic spline.





We compute the test error and obtain 1.462. This is almost the same from what we obtained using the cubic spline. The conclusion is that the cubic spline is as stable as the natural cubic spline in the boundaries. In addition, the larger number of knots being used by the natural cubic spline is not providing any additional improvement.

```
## [1] "Test MSE"
```

```
## [1] 1.460406
```

Smoothing Spline

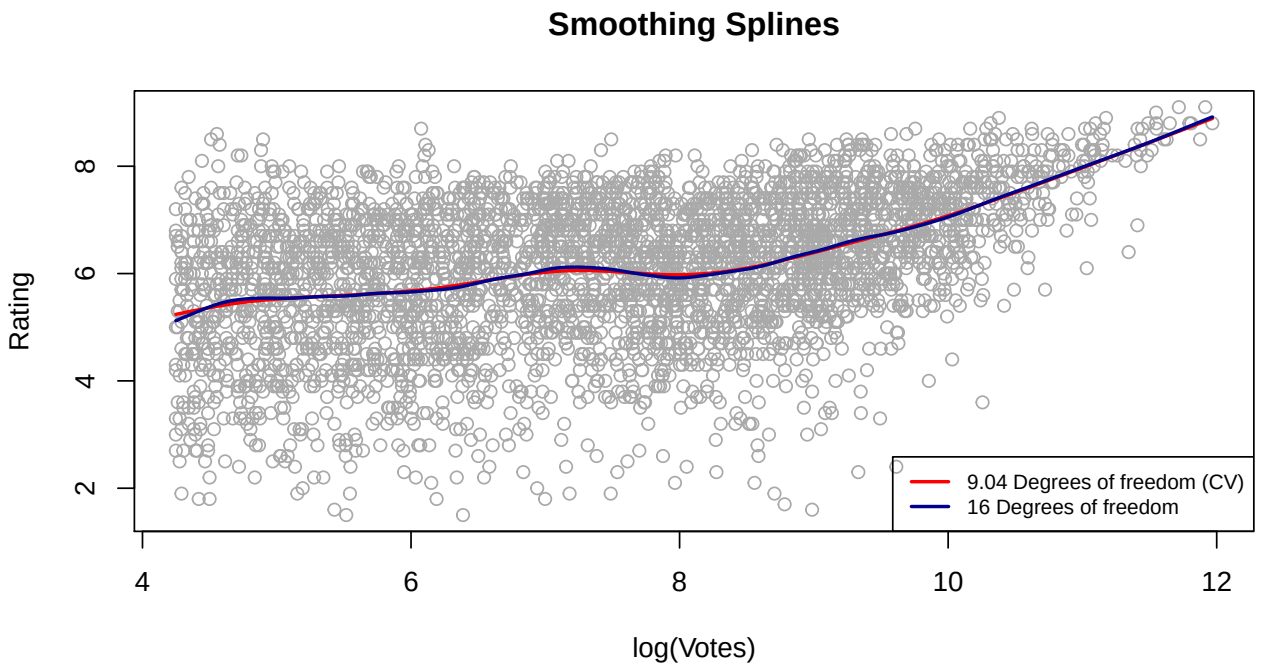
Instead of determining a set of knots and fitting a different function on each one, we try to estimate a function g(x), which provides us with the desired flexibility. The function g(x) should fit the data well, minimizing the RSS,

but nevertheless shouldn't be too rough. Therefore, we introduce a penalty term to control for how wiggly g(x) is. Finally, we tune the penalty using parameter lambda. Higher values of lambda force our function g(x) to be less flexible,

We start by using cross validation to obtain the lambda parameter which minimize the test error estimate. We obtain lambda = 9.047881.

```
## [1] "lambda C.V."
```

```
## [1] 9.047881
```



We compute the test error and obtain 1.457. This model is obtaining a slightly better performance compared to the cubic and natural spline, but not compared to the polynomial regression with the best subset predictors.

```
## [1] "Test MSE"
```

```
## [1] 1.457097
```

General Additive Models (GAMs)

We will now explore General Additive Models (GAMs). These extend the multiple linear regression model by allowing non-linear functions of each variable. We calculate a different function for each variable, and add together all of their contributions. This results in potentially more accurate predictions for our response variable, the movie rating. Furthermore, since the model is additive, we can interpret the effects of each variable on the movie rating.

GAM with Smoothing Spline We use the predictors that were chosen in the best subset selection: year, length, votes, Animation, Comedy, Drama, and adjBudget. We use the 'gam' package in R to fit the models.

```
## Analysis of Deviance Table
##
```

```
## Model 1: rating ~ year + length + votes + Animation + Comedy + Drama +
##     adjBudget
## Model 2: rating ~ s(year, df = 4) + length + votes + Animation + Comedy +
##     Drama + adjBudget
## Model 3: rating ~ s(year, df = 4) + s(length, df = 4) + votes + Animation +
##     Comedy + Drama + adjBudget
## Model 4: rating ~ s(year, df = 4) + s(length, df = 4) + s(votes, df = 4) +
##     Animation + Comedy + Drama + adjBudget
## Model 5: rating ~ s(year, df = 4) + s(length, df = 4) + s(votes, df = 4) +
##     Animation + Comedy + Drama + s(adjBudget, df = 4)
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1      3861     4189.4
## 2      3858     4116.1 3.0001    73.30 < 2.2e-16 ***
## 3      3855     3965.3 3.0003   150.78 < 2.2e-16 ***
## 4      3852     3516.0 2.9999   449.32 < 2.2e-16 ***
## 5      3849     3456.1 3.0003    59.93 2.131e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA results, there is evidence to reject the null hypothesis that the simplest model is sufficient. Thus, the GAM with a smoothing spline for each of year, length, and budget, along with linear variables animation, comedy, and drama, is the best model.

We now make predictions on the training set using the best GAM and compute the test MSE.

```
## [1] 0.9437638
```

The test MSE is 0.944.

GAM with Natural Splines We repeat the process of selecting the best GAM, this time using natural splines instead of smoothing splines. Again, we use the same variables from the best subset selection.

```
## Analysis of Deviance Table
##
## Model 1: rating ~ year + length + votes + Animation + Comedy + Drama +
##     adjBudget
## Model 2: rating ~ ns(year, df = 4) + length + votes + Animation + Comedy +
##     Drama + adjBudget
## Model 3: rating ~ ns(year, df = 4) + ns(length, df = 4) + votes + Animation +
##     Comedy + Drama + adjBudget
## Model 4: rating ~ ns(year, df = 4) + ns(length, df = 4) + ns(votes, df = 4) +
##     Animation + Comedy + Drama + adjBudget
## Model 5: rating ~ ns(year, df = 4) + ns(length, df = 4) + ns(votes, df = 4) +
##     Animation + Comedy + Drama + ns(adjBudget, df = 4)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3861     4189.4
## 2      3858     4118.6  3    70.86 < 2.2e-16 ***
## 3      3855     3969.9  3   148.65 < 2.2e-16 ***
## 4      3852     3400.8  3   569.15 < 2.2e-16 ***
## 5      3849     3303.8  3    97.01 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.974466
```

The natural splines GAM yields a Test MSE of 0.974.

Thus, the best GAM consists of smoothing splines instead of natural splines. With a test MSE of 0.944, this is the most accurate model at this point in our analysis.

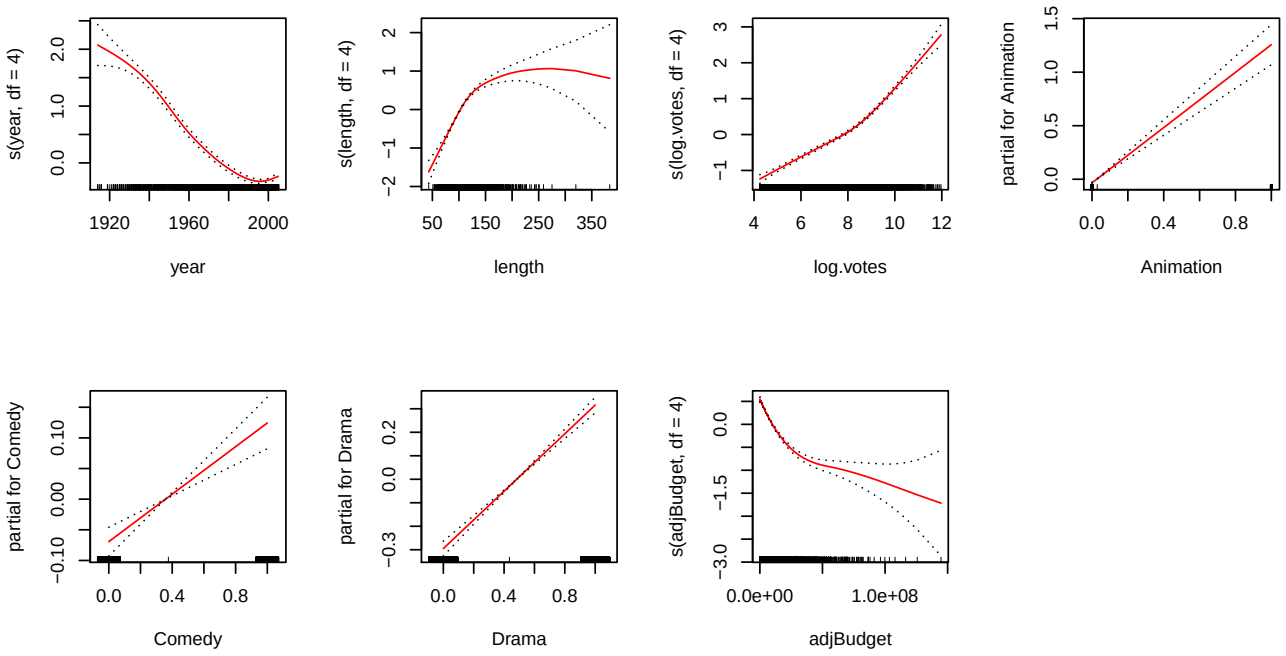
However, we found in our modeling with polynomial regression and splines that taking the logarithm of the votes improves the accuracy of the model. Is this true for the GAM as well? We repeat the analysis above, substituting the logarithm of the votes for the untransformed votes.

GAM with Smoothing Splines and Logarithm of Votes

```
## Analysis of Deviance Table
##
## Model 1: rating ~ year + length + log.votes + Animation + Comedy + Drama +
##     adjBudget
## Model 2: rating ~ s(year, df = 4) + length + log.votes + Animation + Comedy +
##     Drama + adjBudget
## Model 3: rating ~ s(year, df = 4) + s(length, df = 4) + log.votes + Animation +
##     Comedy + Drama + adjBudget
## Model 4: rating ~ s(year, df = 4) + s(length, df = 4) + s(log.votes, df = 4) +
##     Animation + Comedy + Drama + adjBudget
## Model 5: rating ~ s(year, df = 4) + s(length, df = 4) + s(log.votes, df = 4) +
##     Animation + Comedy + Drama + s(adjBudget, df = 4)
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1      3861     3684.0
## 2      3858     3561.6 3.0001  122.399 < 2.2e-16 ***
## 3      3855     3503.5 3.0003   58.066 1.021e-14 ***
## 4      3852     3376.7 2.9999  126.844 < 2.2e-16 ***
## 5      3849     3275.5 3.0003  101.179 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.9367831
```

Based on the ANOVA table, the best GAM that uses smoothing splines and the logarithmic votes is the most complex, with a spline for year, length, log(votes), and adjBudget. It has a test MSE of 0.937. This is slightly better than the test MSE for the identical model with normal votes instead of log(votes), which had a test MSE of 0.944.



The plot of our GAM displays a plot for each variable, showing its individual contribution to the additive model. The standard errors are represented by the dotted lines.

GAM with Natural Splines and Logarithm of Votes

```
## Analysis of Deviance Table
##
## Model 1: rating ~ year + length + log.votes + Animation + Comedy + Drama +
##     adjBudget
## Model 2: rating ~ ns(year, df = 4) + length + log.votes + Animation +
##     Comedy + Drama + adjBudget
## Model 3: rating ~ ns(year, df = 4) + ns(length, df = 4) + log.votes +
##     Animation + Comedy + Drama + adjBudget
## Model 4: rating ~ ns(year, df = 4) + ns(length, df = 4) + ns(log.votes,
##     df = 4) + Animation + Comedy + Drama + adjBudget
## Model 5: rating ~ ns(year, df = 4) + ns(length, df = 4) + ns(log.votes,
##     df = 4) + Animation + Comedy + Drama + ns(adjBudget, df = 4)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3861     3684.0
## 2      3858     3563.6  3  120.434 < 2.2e-16 ***
## 3      3855     3504.8  3   58.726  6.94e-15 ***
## 4      3852     3376.2  3  128.624 < 2.2e-16 ***
## 5      3849     3275.2  3  101.012 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.9798634
```

The best GAM that uses natural splines and logarithmic votes has a test MSE of 0.980, making it the least accurate of the GAMs.

We conclude that the GAM with smoothing splines is the best model for predicting the rating of a movie. Furthermore, taking the logarithm of the votes provides a slight improvement to the accuracy of the model; however, the benefit of the logarithm was more substantial in our linear and polynomial regression models.

Conclusion

Now that we have performed exploratory analysis on our variables and fitted a number of linear and non-linear models to the data, we can derive some observations from it. First, below we present a summary of the fits on different models, sorted by ascending predicting power (Test MSE).

Model	Variables	Test MSE	R-sq
Spline	lg(*votes*)	1.464	0.192
Natural Splines	lg(*votes*)	1.460	0.192
Polynomial D3 Regression	lg(*votes*)	1.460	0.187
Smoothing Splines	lg(*votes*)	1.457	*n/a*
Linear Regression	*year, length, votes, Animation, Comedy, Drama, adjBudget*	1.089	0.396
Linear Regression	*year, length,* lg(*votes*)*, Animation, Comedy, Drama, adjBudget*	1.037	0.479
Polynomial D3 Regression	*year, length,* lg(*votes*)*, Animation,Comedy, Drama, adjBudget*	0.991	0.474
GAM (N. Spline)	*year, length,* lg(*votes*)*, Animation, Comedy, Drama, adjBudget*	0.979	*n/a*
GAM (N. Spline)	*year, length, votes, Animation, Comedy, Drama, adjBudget*	0.974	*n/a*
GAM (S. Splines)	*year, length, votes, Animation, Comedy, Drama, adjBudget*	0.943	*n/a*
GAM (S. Splines)	*year, length,* lg(*votes*)*, Animation, Comedy, Drama, adjBudget*	0.937	*n/a*

Throughout this exercise we identified *votes* as the regressor with the greatest exploratory power on any given model, something we noted when we analyzed the correlation matrix. This variable alone, when regressed upon on the Smoothing Splines model, achieved a MSE test error of 1.456 and an R^2 of almost 20%, which is relatively high considering the complexity of the dataset.

If we incorporate more variables into our models, specifically those identified by the Best Subset Selection technique, we are able to enhance our predicting power significantly. This technique suggested utilizing a relatively large number of variables (7), which implies that our main concern with this dataset related to the *Bias* component of the MSE.

On the multivariate linear regression and polynomial models, we see the following direction on their coefficients:

year	(−)	Animation	(+)
length	(+)	Drama	(+)
votes	(+)	adjBudget	(−)
Comedy	(+)		

In terms of interpretability, some of these signs makes sense and are consistent with what we saw on the correlation matrix: year is negative, while length, votes and Drama are positive. The signs for the other 3 regressors are the opposite ones of the observed on the correlation matrix, but this needs not to be contradictory as our models take into account how all these variables come into play. The GAM and splines models resulted harder to interpret, and we see GAMs performing slightly better than our Polynomial D3 regression.

If we are to favor the *prediction* capabilities of our models, one would think of using the GAM (Smoothed Splines), which combines the power of non-parametric and linear models to provide greater flexibility. However, one should also note that the polynomial model is almost as good, and its *intepretation* is easier, so we would recommend sticking to the multivariate polynomial regression in this case.

For future studies we would recommend incorporating classification techniques into the analysis, as this distinct approach might yield interesting results (while treating rating as a categorical variable, rather than a continuous one).