Analysis and Prediction on Movies

Final Project - Data Visualization and Analytics

Desared Osmanllari

10.02.2017

1. Analysis of the relationship among movie variables

1.1 Data

The data comes from the Rotten Tomatoes and IMDB web sites and consists of audience and critic review scores for 651 movies (random sample of data) released from 1970 till 2016. In addition to review scores, the data contains several other variables regarding each movie such as genre, running time, MPAA rating, production studio, Oscar nominations, release year, actors participating and more. In total, there is an amount of 32 variables stored in this dataset.

The raw data is a random sample taken from the full data set of all movies released prior to 2016. It is assumed for the sake of exploratory analysis that the conclusions drawn are generalizable to the population of all movies and that there is no bias introduced by the sampling method. But this is supposed just for the exploratory analysis phase, which will help us to create a general idea about the dataset. After drawing the conclusions from exploratory analysis, I will split the data into two subset: training and testing set.

I will start this project by introducing the data, showing their correlations and deciding wich of the variables is the most appropriate one to be furtherly studied as a dependent variable. Later, I will make predictions using various regression methods and show how some interesting components affect the movie rating.

Exploratory analysis and visualizations are located in the Appendix to this document.

1.2 Setup

Load packages and data

```
# set working directory
setwd("C:/RWTH/ws16/Data Visualisation and Analytics/Final Project")
#load all necessary packages
require(pacman)
p_load( ggplot2, dplyr,statsr,grid,gridExtra, knitr, corrplot,
        corrr, ggthemes, plotly ,data.table, formattable , DT, glmnet,
        caret, leaps,e1071, GGally, car) #load the dataset
load("movies.Rdata")
# show the features of the dataset
dim(movies)
```

```
## [1] 651  32
```

```
str(movies)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    651 obs. of  32 variables:
##  $ title           : chr  "Filly Brown" "The Dish" "Waiting for Guffman" "The Age of Innocence" ...
##  $ title_type      : Factor w/ 3 levels "Documentary",..: 2 2 2 2 2 1 2 2 1 2 ...
##  $ genre           : Factor w/ 11 levels "Action & Adventure",..: 6 6 4 6 7 5 6 6 5 6 ...
##  $ runtime         : num  80 101 84 139 90 78 142 93 88 119 ...
```

```
##  $ mpaa_rating     : Factor w/ 6 levels "G","NC-17","PG",..: 5 4 5 3 5 6 4 5 6 6 ...
##  $ studio          : Factor w/ 211 levels "20th Century Fox",..: 91 202 167 34 13 163 147 118 88 84
##  $ thtr_rel_year   : num  2013 2001 1996 1993 2004 ...
##  $ thtr_rel_month  : num  4 3 8 10 9 1 1 11 9 3 ...
##  $ thtr_rel_day    : num  19 14 21 1 10 15 1 8 7 2 ...
##  $ dvd_rel_year    : num  2013 2001 2001 2001 2005 ...
##  $ dvd_rel_month   : num  7 8 8 11 4 4 2 3 1 8 ...
##  $ dvd_rel_day     : num  30 28 21 6 19 20 18 2 21 14 ...
##  $ imdb_rating     : num  5.5 7.3 7.6 7.2 5.1 7.8 7.2 5.5 7.5 6.6 ...
##  $ imdb_num_votes  : int  899 12285 22381 35096 2386 333 5016 2272 880 12496 ...
##  $ critics_rating  : Factor w/ 3 levels "Certified Fresh",..: 3 1 1 1 3 2 3 3 2 1 ...
##  $ critics_score   : num  45 96 91 80 33 91 57 17 90 83 ...
##  $ audience_rating : Factor w/ 2 levels "Spilled","Upright": 2 2 2 2 1 2 2 1 2 2 ...
##  $ audience_score  : num  73 81 91 76 27 86 76 47 89 66 ...
##  $ best_pic_nom    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ best_pic_win    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ best_actor_win  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 2 1 1 ...
##  $ best_actress_win: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ best_dir_win    : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ top200_box      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ director        : chr  "Michael D. Olmos" "Rob Sitch" "Christopher Guest" "Martin Scorsese" ...
##  $ actor1          : chr  "Gina Rodriguez" "Sam Neill" "Christopher Guest" "Daniel Day-Lewis" ...
##  $ actor2          : chr  "Jenni Rivera" "Kevin Harrington" "Catherine O'Hara" "Michelle Pfeiffer" .
##  $ actor3          : chr  "Lou Diamond Phillips" "Patrick Warburton" "Parker Posey" "Winona Ryder" .
##  $ actor4          : chr  "Emilio Rivera" "Tom Long" "Eugene Levy" "Richard E. Grant" ...
##  $ actor5          : chr  "Joseph Julian Soria" "Genevieve Mooy" "Bob Balaban" "Alec McCowen" ...
##  $ imdb_url        : chr  "http://www.imdb.com/title/tt1869425/" "http://www.imdb.com/title/tt020587
##  $ rt_url          : chr  "//www.rottentomatoes.com/m/filly_brown_2012/" "//www.rottentomatoes.com/m/
```

Research Question

The research question addressed in this analysis is this:

- Is there any correlations between movie general features that make people like or dislike a movie?

More specifically, do variables such as movie genre, MPAA rating, run length, etc. work as reasonable predictors of a movie's popularity? Can we predict the rating of a movie based on these components?

Predicting which movies are likely to be popular ahead of their release would be valuable information when deciding which movies to watch. Moreover, companies such as Netflix do such an analysis to predict wich movies are mostly preferable in specific seasons.

1.3 General Exploratory Analysis

What are the total number of movies reviewed by year?

From the first graph (Appendix - Plot 1), I show that the total number of movies produced in the last 40 year keeps increasing. The trend is visualized with the blue trace. It does not mean that the number of movies per year is decreasing in the last decade. The sample might show unacurrate correlations, but in overall we can see a positive correlation among years and movies produced. This is somehow predictable, since the technology and budgets for producing new movies keeps increasing every year. This makes the last decade the most successful one in the world of cinematography.

Data Characteristics

As noted before, there are 651 movies represented in the raw data. The charts (Appendix - Plot 2) show a breakdown of the type of movies included in the sample. I have visualized the distribution of movies in 4 main variables. From this chart, I can better understand the dataset. I can conclude that most of the movies belong to 'Feature Film' type and 'Drama' genre. Also, most of the movies have a MPAA rating of 'R'. Movie runtime is normally distributed and we see that most of the movies have an average runtime of 100 minutes.

Correlation of all numeric variables.

Next I study the correlation of all numeric variables (Appendix - Plot 3). First I clean the dataset from null values and estimate the correlated variables. This feature is important to decide the significance that numeric variables have to each other. From the correlation plot, I can say that the most correlated variables are audience_score and imdb_rating. These two variables are related with critics_score also. Moreover, dvd_release_year is correlated to theater_release_year, which makes sense since the dvd is usually released some months later than the movie is showed in cinemas.

1.4 Modeling

Model Development

The target response variable for the prediction model is a movie rating score, but with three to choose from, which one should be used?

(1) Average of reviews by movie critics .

(2) Average of reviews from the public (a.k.a., audience).

(3) Average of reviews on the IMDB web site (no distinction made between critics and audience reviews).

For this reason, I prepare a correlation between the different rating scores. The plots in Appendix - Plot 4 show that to be the case.

Given these correlations, only one of the ratings will be used as the response variable. Histograms of the data distribution for the three ratings are given in Appendix - Plot 5.

Contrary to the two Rotten Tomatoe scores, the IMDB scores show a nice, mostly normal distribution centered around a mean of 6.37 with somewhat of a left-side skew. Given its distribution and the fact that it has the highest pairwise correlation with the other scores, the IMDB rating (imdb_rating) was the chosen response variable.

1.5 Further Exploratory Analysis considering IMDB rating

(a) Analyse correlation of IMDB rating with other variables

In the graph displayed in Appendix - Plot 6, I have removed removed the critics and audience score, since it is proved they have the highest correlation value with the IMDB rating. My purpose is to see which variables affect mostly the IMDB. As visualized, movie runtime and the number of IMDB votes have the highest values among the other numeric values. Later I will make a more detailed analysis and prediction based on these indepentend variables.

(b) Number of users voting versus IMDB score

In Appendix - Plot 7, there is a positive curvilinear relationship among imdb_num_votes and imdb_rating. This relation is somehow predictable since as many users vote for a specific movie, the higher is the possibility for this movie to have a good rating. With red line, there is visualized the Fitting Linear Model.

(c) Movie runtime versus IMDB score

Even though, the correlation among these two variables is the second highest, it is hard to say from the graph that runtime affects the imdb_rating. This is visualized in Appendix - Plot 8. Actually, runtime within a specific interval really affects the imdb_rating. So we have to consider an interval while studying this correlation, since it is hard to believe that a 5-hour longtime movie has a higher rating than a normal 90 minutes movie.

(d) Content rating versus IMDB rating

From the graph in Appendix - Plot 9, we can assume that movies belonging to MPAA rating "G", are tended to have a higher IMDB compared to the other ratings. However, the sample lacks some information. So ,it is hard to say whether mpaa rating 'NC-17' has an effect on imdb rating.

(e) IMDB rating distribution over years

Not concrete results can be implied by the graph in Appendix - Plot 10. The distribution of movies through the years changes as seen previously, but it does not affect the imdb rating. Yet we can not be sure only from the visualization without checking their regression relationship.

(f) Which director has the highest average IMBD rating?

From Appendix - Plot 11, I can say that directors might affect the quality of a movie considerably. Consequently, they might be a decisive factor on imdb_rating. As seen in the barplot, there is a wide distribution of average imdb_rating per director.

1.6 Data Manipulation

After doing some explonatory analysis and better understanding the data, I try doing some predictions. However, before making the first viewings and processings over the data, we split our dataset in training and test data. Most of the algorithms that learn with data with the objective to predict new incomes try to minimize the error they make with the first step (learning). Because of this, if we evaluate our data over the error that the algorithm (regression in this case) makes with the training data, we would be making an over optimistic assumption. Then, we make a validation using the testing set to see how realistic our prediction is and check whether our model prediction responds similarly for both training and testing set.

```
set.seed(123)
anyNA <- function(row){ any(is.na(row))}      # function to return which row has at least a NA
rowsWithNA <- apply(movies, 1, anyNA)         # apply the function to each row
sum(rowsWithNA)
```

```
## [1] 32
```

```
movies <- movies[!rowsWithNA,]

responseCols <- c(13)                         #imdb rating is dependent variable
 # remove title, critics_rating,critics_score,audience_rating,audience_score
# remove the name of directors and actors and the url of imdb and rotten tomatos
```

```
noInterestVars <- c(1,15,16,17,18,25:32)
inTrain <- createDataPartition(y=movies$imdb_rating,p=0.6, list=FALSE)
training <- movies[inTrain,-noInterestVars]          # create training set
testing <- movies[-inTrain,-noInterestVars]          # create testing set
dim(training); dim(testing)
```

```
## [1] 374  19
```

```
## [1] 245  19
```

2. Prediction on Movies

2.1 Model 1 - Regression on multiple covariates

The purpose of multiple regression is to predict a single variable from one or more independent variables. Multiple regression with many predictor variables is an extension of linear regression with two predictor variables. A linear transformation of the X variables is done so that the sum of squared deviations of the observed and predicted Y is a minimum.

Basic idea

(1) Fit a regression model
(2) Penalize (or shrink) large coefficients

Pros:

- Can help with the bias/variance tradeoff
- Can help with model selection

Cons:

- May be computationally demanding on large data sets
- Does not perform as well as random forests and boosting

Model selection approach: split samples

Approach:

(1) Divide data into training/test/validation
(2) Treat validation as test data, train all competing models on the train data and pick the best one on validation.

(3) To appropriately assess performance on new data apply to test set
(4) You may re-split and reperform steps 1-3

Two common problems:

(a) Limited data
(b) Computational complexity

A motivating example

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where X_1 and X_2 are nearly perfectly correlated (co-linear). You can approximate this model by:

$$Y = \beta_0 + (\beta_1 + \beta_2) X_1 + \epsilon$$

The result is:

- You will get a good estimate of Y
- The estimate (of Y) will be biased
- We may reduce variance in the estimate

Adjusting Explanatory Variables

Criteria for model selection:

(1) The data are normal distributed

(2) R^2 always increase as parameters are added

(3) Adjusted R^2: generally favors models with too many variables

(4) F-test: statistical test for normal, nested models.

More general criteria:

(1) Akaike's information criterion (AIC):

$$nlog(\hat{\sigma}^2) + 2p$$

Penalizes the number of parameters or coeffiecients in the model. The Akaike Information Criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

(2) Bayesian information criterion (BIC):

$$nlog(\hat{\sigma}^2) + log(n)p$$

(3) Cross Validation

Models which keep all the variables but shrinking them toward zero:

(1) Lasso

$\sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$ subject to $\sum_{j=1}^{p} |\beta_j| \leq s$

also has a lagrangian form

$$\sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

(2) Ridge Regression

$$\sum_{i=1}^{N}\left(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

equivalent to solving

$\sum_{i=1}^{N}\left(y_i-\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$ subject to $\sum_{j=1}^{p}\beta_j^2\leq s$ where s is inversely proportional to λ

What does λ shrink the model?

(a) λ controls the size of the coefficients
(b) λ controls the amount of regularization
(c) As $\lambda\to 0$ we obtain the least square solution
(d) As $\lambda\to\infty$ we have $\hat{\beta}_{\lambda=\infty}^{ridge}=0$

Implementing Akaike's Information criterion (AIC)

I decided to use AIC selection model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, because increasing the number of parameters in the model almost always improves the calculated goodness of the fit.

I perform a stepwise variable selection procedure. I used backward selection model which basically starts from a full model and removes less significant explanatory variables.

This model removes one variable at a time and calculates the value of AIC in each step. I will try to get the smallest value of AIC. After reaching the best option, I stop removing variables.

```
# make a linear model regression on multiple covariance: imdb_rating as dependent variable
# and all the others as independent
set.seed(123)
modFitAll <- lm(imdb_rating ~ .,data = training)
summary(modFitAll)
```

```
##
## Call:
## lm(formula = imdb_rating ~ ., data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2659 -0.2172  0.0000  0.2179  1.8496
##
## Coefficients: (2 not defined because of singularities)
##                                              Estimate Std. Error
## (Intercept)                                 7.869e+01  3.424e+01
## title_typeFeature Film                     -1.386e+00  5.335e-01
## title_typeTV Movie                          7.738e-01  1.071e+00
## genreAnimation                              8.969e-01  5.647e-01
## genreArt House & International              1.201e+00  5.205e-01
## genreComedy                                 2.094e-01  2.270e-01
## genreDocumentary                            5.819e-01  5.780e-01
## genreDrama                                  7.841e-01  1.935e-01
## genreHorror                                -1.717e-02  3.585e-01
## genreMusical & Performing Arts              1.164e+00  4.184e-01
```

```
## genreMystery & Suspense                          5.065e-01  2.546e-01
## genreOther                                       4.337e-01  4.320e-01
## genreScience Fiction & Fantasy                   2.892e-02  6.235e-01
## runtime                                          3.756e-03  4.047e-03
## mpaa_ratingNC-17                                 3.126e-02  1.066e+00
## mpaa_ratingPG                                   -1.040e+00  5.063e-01
## mpaa_ratingPG-13                                -1.238e+00  5.459e-01
## mpaa_ratingR                                    -1.003e+00  5.399e-01
## mpaa_ratingUnrated                              -2.375e-01  9.028e-01
## studio20th Century Fox Film Corporat             5.814e-01  9.614e-01
## studio20th Century Fox Film Corporation          1.247e+00  6.317e-01
## studio7-57 Releasing                             1.065e+00  1.052e+00
## studio905 Corporation                            5.547e-01  9.327e-01
## studioA24                                        1.263e+00  9.610e-01
## studioA24 Films                                  1.030e+00  7.407e-01
## studioAlliance Atlantis Communications           1.233e+00  9.275e-01
## studioAnalysis                                   9.474e-01  9.485e-01
## studioAnchor Bay Films                           1.951e+00  9.838e-01
## studioArtisan Entertainment                      1.595e+00  9.339e-01
## studioAVCO Embassy Pictures                      1.341e+00  9.701e-01
## studioBankside Films                            -6.009e-01  1.359e+00
## studioBrainstorm Media                           1.616e+00  1.190e+00
## studioBuena Vista                                4.833e-01  6.259e-01
## studioBuena Vista Internationa                   1.722e+00  9.283e-01
## studioBuena Vista Pictures                       2.047e-01  6.031e-01
## studioCarnaby International                      2.355e+00  9.344e-01
## studioCine-Source                                7.262e-01  9.464e-01
## studioCinema Libre Studio                        1.322e+00  1.249e+00
## studioCinema Seven Productions Ltd               1.340e+00  9.325e-01
## studioCode Red                                   9.231e-01  1.282e+00
## studioColumbia Pictures                          8.892e-01  6.359e-01
## studioColumbia Tristar Pictures                  1.353e+00  9.183e-01
## studioConcorde/New Horizons Home Video          -1.206e+00  9.704e-01
## studioCowboy Pictures                            1.118e+00  8.346e-01
## studioCriterion Collection                       1.404e+00  1.007e+00
## studioCrown International Pictures              -1.413e+00  1.104e+00
## studioDestination Films                         -1.924e+00  9.287e-01
## studioDiva                                       3.598e-01  1.216e+00
## studioDreamworks                                 4.665e-01  9.575e-01
## studioDreamWorks Studios                         9.207e-01  1.001e+00
## studioE1 Entertainment                          -3.405e+00  1.368e+00
## studioEcho Bridge Home Entertainment             1.174e+00  9.383e-01
## studioEmbassy                                    2.258e+00  9.508e-01
## studioFabrication Films                          2.145e+00  9.316e-01
## studioFilm Movement                              1.699e+00  1.280e+00
## studioFilmDistrict                               1.907e+00  9.239e-01
## studioFirst Run Entertainment                    8.864e-01  1.219e+00
## studioFirst Run Features                         4.785e-01  1.001e+00
## studioFocus Features                             9.127e-01  6.462e-01
## studioFox Atomic                                -1.079e-01  9.408e-01
## studioFox Searchlight Pictures                   2.493e+00  9.503e-01
## studioFreestyle Releasing                       -4.713e-01  9.421e-01
## studioGenius Productions                         8.198e-01  9.482e-01
## studioGood Machine                               1.696e+00  9.373e-01
```

```
## studioGravitas                                1.101e+00  9.275e-01
## studioGreyCat Films                                  NA         NA
## studioGroup 1                                 1.138e+00  1.103e+00
## studioHatchet Films                           8.730e-01  9.307e-01
## studioHBO Video                                      NA         NA
## studioHemdale                                 1.584e+00  9.399e-01
## studioHK Film Corporation                    -6.187e-01  9.306e-01
## studioHollywood Pictures                     -7.981e-01  9.276e-01
## studioIcarus Films                            7.218e-01  1.210e+00
## studioIFC                                    -6.517e-02  1.185e+00
## studioIFC Films                               1.307e+00  5.953e-01
## studioIFC First Take                          2.372e+00  9.355e-01
## studioIFC Midnight                            1.058e+00  1.165e+00
## studioImage Entertainment                     1.359e+00  7.796e-01
## studioIndependent Pictures                    6.514e-01  1.079e+00
## studioIndomina Films                          1.424e+00  1.006e+00
## studioIndomina Media Inc.                     7.220e-01  9.447e-01
## studioKaga Bay                                9.645e-01  1.004e+00
## studioLions Gate Films                        1.602e+00  6.123e-01
## studioLions Gate Films Inc.                   1.379e+00  9.669e-01
## studioLions Gate Releasing                    1.279e+00  7.481e-01
## studioLionsgate                              -1.669e+00  7.324e-01
## studioLionsgate Films                         7.802e-01  7.486e-01
## studioLionsgate Releasing                     1.458e+00  9.242e-01
## studioLive Home Video                         1.317e+00  9.317e-01
## studioLorimar Home Video                      2.218e+00  9.524e-01
## studioMadman Entertainment                    1.430e+00  9.925e-01
## studioMagic Lamp Releasing                    6.039e-01  9.342e-01
## studioMagnet Releasing                        1.582e+00  9.370e-01
## studioMagnet/Magnolia Pictures                2.178e+00  9.315e-01
## studioMagnolia Pictures                       1.493e+00  5.589e-01
## studioMCA Universal Home Video                7.037e-01  5.814e-01
## studioMetro-Goldwyn-Mayer Pictures            1.263e+00  1.018e+00
## studioMGM                                     5.695e-01  5.112e-01
## studioMGM Home Entertainment                  7.367e-01  5.792e-01
## studioMiramax                                 1.063e+00  5.635e-01
## studioMiramax Films                           9.976e-01  5.027e-01
## studioMusic Box Films                         2.568e+00  9.427e-01
## studioNelson Entertainment                    3.575e-01  7.545e-01
## studioNew Line Cinema                         7.260e-01  5.293e-01
## studioNew Line Home Entertainment             7.113e-01  5.907e-01
## studioNew World Pictures                      1.678e+00  9.541e-01
## studioNew Yorker Films                        1.556e+00  9.559e-01
## studioNewmarket Film Group                    2.081e+00  9.422e-01
## studioNewmarket Films                        -5.669e-01  1.060e+00
## studioNordisk Film Biograf Distribution       1.490e+00  1.189e+00
## studioNordisk Film Biografdistributi          2.115e+00  9.351e-01
## studioOctober Films                           1.682e+00  9.554e-01
## studioOpen Road Films                         4.204e-01  9.450e-01
## studioOrion Home Video                        1.096e+00  6.592e-01
## studioOrion Pictures Corporation              3.530e-01  7.324e-01
## studioOutsider Films                          2.455e+00  9.476e-01
## studioOverture Films                          2.268e+00  9.336e-01
## studioParamount                              -6.538e-01  6.691e-01
```

```
## studioParamount Home Video                        1.002e+00  5.572e-01
## studioParamount Pictures                           8.390e-01  4.617e-01
## studioParamount Studios                            1.669e+00  7.386e-01
## studioRelativity Media                             1.244e+00  9.458e-01
## studioRoadside Attractions                         1.743e+00  7.673e-01
## studioSag Harbor-Basement Pictures                 8.712e-01  1.210e+00
## studioSaguenay Films                               1.337e-01  1.241e+00
## studioSamuel Goldwyn Films                         1.723e+00  9.539e-01
## studioScreen Gems                                  6.796e-01  9.426e-01
## studioSeventh Art Productions                     -7.156e-02  1.310e+00
## studioSony Pictures                                8.684e-01  5.297e-01
## studioSony Pictures Classics                       1.294e+00  5.884e-01
## studioSony Pictures Entertainment                  9.342e-01  6.674e-01
## studioSony Pictures Home Entertainment             7.662e-01  4.917e-01
## studioSony Pictures/Columbia                       1.399e+00  9.368e-01
## studioSony Pictures/Screen Gems                    2.306e+00  9.410e-01
## studioStrand Releasing                             9.788e-01  1.065e+00
## studioSummit Entertainment                         2.335e+00  7.903e-01
## studioTango Entertainment                          7.570e-01  9.586e-01
## studioThe Film Arcade                              1.543e+00  9.507e-01
## studioThe Shooting Gallery                         2.191e+00  9.462e-01
## studioThe Weinstein Co.                            1.563e+00  7.292e-01
## studioThe Weinstein Company                        1.846e+00  7.035e-01
## studioThinkFilm                                    1.336e+00  9.500e-01
## studioTouchstone Home Entertainment                1.640e+00  9.501e-01
## studioTouchstone Pictures                          1.404e+00  6.518e-01
## studioTrimark                                      1.493e+00  7.329e-01
## studioTriStar                                      1.882e+00  9.359e-01
## studioTriStar Pictures                             1.746e-02  1.004e+00
## studioTwentieth Century Fox Home Entertainment     9.685e-01  5.349e-01
## studioUnited Artists                               2.030e+00  6.544e-01
## studioUniversal Pictures                           1.338e+00  4.699e-01
## studioUniversal Studios                            1.134e+00  7.269e-01
## studioUrban Vision Entertainment                   3.181e+00  9.243e-01
## studioUSA Films                                    1.280e+00  9.390e-01
## studioVirgin Vision                                9.483e-01  9.395e-01
## studioWalt Disney Pictures                         1.251e+00  9.386e-01
## studioWalt Disney Productions                      8.451e-01  9.402e-01
## studioWarner Bros Pictures                         6.439e-01  9.454e-01
## studioWarner Bros.                                 1.020e+00  5.745e-01
## studioWarner Bros. Pictures                        1.201e+00  4.579e-01
## studioWARNER BROTHERS PICTURES                     2.231e-01  6.107e-01
## studioWarner Home Video                            7.937e-01  5.070e-01
## studioWarner Independent                           1.769e+00  9.311e-01
## studioWarner Independent Pictures                  1.300e+00  9.233e-01
## studioWarners Bros. Pictures                       1.467e+00  9.602e-01
## studioWeinstein Company                            1.526e+00  9.278e-01
## studioWinstar                                      1.930e+00  9.329e-01
## studioYari Film Group Releasing                    1.838e+00  9.331e-01
## studioZeitgeist Films                              5.472e-01  1.212e+00
## thtr_rel_year                                     -2.290e-02  9.106e-03
## thtr_rel_month                                     1.981e-02  1.831e-02
## thtr_rel_day                                       9.352e-03  6.752e-03
## dvd_rel_year                                      -1.337e-02  1.935e-02
```

```
## dvd_rel_month                                2.384e-02  1.724e-02
## dvd_rel_day                                 -7.562e-04  6.351e-03
## imdb_num_votes                               4.899e-06  7.362e-07
## best_pic_nomyes                              7.009e-01  3.491e-01
## best_pic_winyes                             -2.055e+00  6.084e-01
## best_actor_winyes                           -1.512e-01  1.677e-01
## best_actress_winyes                         -3.341e-01  1.732e-01
## best_dir_winyes                              4.763e-01  1.974e-01
## top200_boxyes                               -1.051e-01  3.446e-01
##                                             t value Pr(>|t|)
## (Intercept)                                   2.298 0.022617 *
## title_typeFeature Film                       -2.597 0.010119 *
## title_typeTV Movie                            0.723 0.470728
## genreAnimation                                1.588 0.113839
## genreArt House & International                2.308 0.022031 *
## genreComedy                                   0.922 0.357418
## genreDocumentary                              1.007 0.315291
## genreDrama                                    4.052 7.34e-05 ***
## genreHorror                                  -0.048 0.961848
## genreMusical & Performing Arts                2.782 0.005933 **
## genreMystery & Suspense                       1.989 0.048044 *
## genreOther                                    1.004 0.316656
## genreScience Fiction & Fantasy                0.046 0.963053
## runtime                                       0.928 0.354596
## mpaa_ratingNC-17                              0.029 0.976640
## mpaa_ratingPG                                -2.055 0.041221 *
## mpaa_ratingPG-13                             -2.269 0.024393 *
## mpaa_ratingR                                 -1.858 0.064681 .
## mpaa_ratingUnrated                           -0.263 0.792791
## studio20th Century Fox Film Corporat          0.605 0.546073
## studio20th Century Fox Film Corporation       1.975 0.049698 *
## studio7-57 Releasing                          1.012 0.312938
## studio905 Corporation                         0.595 0.552709
## studioA24                                     1.314 0.190441
## studioA24 Films                               1.391 0.165950
## studioAlliance Atlantis Communications        1.329 0.185284
## studioAnalysis                                0.999 0.319082
## studioAnchor Bay Films                        1.983 0.048811 *
## studioArtisan Entertainment                   1.708 0.089300 .
## studioAVCO Embassy Pictures                   1.382 0.168607
## studioBankside Films                         -0.442 0.658934
## studioBrainstorm Media                        1.358 0.176093
## studioBuena Vista                             0.772 0.440895
## studioBuena Vista Internationa                1.854 0.065195 .
## studioBuena Vista Pictures                    0.339 0.734669
## studioCarnaby International                   2.521 0.012520 *
## studioCine-Source                             0.767 0.443849
## studioCinema Libre Studio                     1.058 0.291232
## studioCinema Seven Productions Ltd            1.437 0.152455
## studioCode Red                                0.720 0.472190
## studioColumbia Pictures                       1.398 0.163568
## studioColumbia Tristar Pictures               1.473 0.142360
## studioConcorde/New Horizons Home Video       -1.243 0.215391
## studioCowboy Pictures                         1.339 0.181978
```

```
## studioCriterion Collection                    1.393 0.165110
## studioCrown International Pictures            -1.280 0.202131
## studioDestination Films                       -2.072 0.039561 *
## studioDiva                                     0.296 0.767529
## studioDreamworks                               0.487 0.626685
## studioDreamWorks Studios                       0.920 0.358941
## studioE1 Entertainment                        -2.489 0.013660 *
## studioEcho Bridge Home Entertainment           1.251 0.212410
## studioEmbassy                                  2.375 0.018506 *
## studioFabrication Films                        2.302 0.022370 *
## studioFilm Movement                            1.327 0.186061
## studioFilmDistrict                             2.065 0.040286 *
## studioFirst Run Entertainment                  0.727 0.468020
## studioFirst Run Features                       0.478 0.633261
## studioFocus Features                           1.412 0.159443
## studioFox Atomic                              -0.115 0.908817
## studioFox Searchlight Pictures                 2.624 0.009383 **
## studioFreestyle Releasing                     -0.500 0.617436
## studioGenius Productions                       0.865 0.388338
## studioGood Machine                             1.810 0.071890 .
## studioGravitas                                 1.187 0.236811
## studioGreyCat Films                               NA       NA
## studioGroup 1                                  1.031 0.303780
## studioHatchet Films                            0.938 0.349399
## studioHBO Video                                   NA       NA
## studioHemdale                                  1.686 0.093479 .
## studioHK Film Corporation                     -0.665 0.506973
## studioHollywood Pictures                      -0.860 0.390666
## studioIcarus Films                             0.596 0.551671
## studioIFC                                     -0.055 0.956188
## studioIFC Films                                2.196 0.029284 *
## studioIFC First Take                           2.536 0.012000 *
## studioIFC Midnight                             0.908 0.364984
## studioImage Entertainment                      1.743 0.082831 .
## studioIndependent Pictures                     0.604 0.546796
## studioIndomina Films                           1.416 0.158430
## studioIndomina Media Inc.                      0.764 0.445659
## studioKaga Bay                                 0.961 0.337894
## studioLions Gate Films                         2.617 0.009575 **
## studioLions Gate Films Inc.                    1.426 0.155388
## studioLions Gate Releasing                     1.710 0.088908 .
## studioLionsgate                               -2.278 0.023802 *
## studioLionsgate Films                          1.042 0.298631
## studioLionsgate Releasing                      1.578 0.116299
## studioLive Home Video                          1.414 0.158971
## studioLorimar Home Video                       2.329 0.020865 *
## studioMadman Entertainment                     1.441 0.151328
## studioMagic Lamp Releasing                     0.647 0.518708
## studioMagnet Releasing                         1.688 0.092997 .
## studioMagnet/Magnolia Pictures                 2.338 0.020404 *
## studioMagnolia Pictures                        2.671 0.008192 **
## studioMCA Universal Home Video                 1.210 0.227603
## studioMetro-Goldwyn-Mayer Pictures             1.241 0.216205
## studioMGM                                      1.114 0.266591
```

```
## studioMGM Home Entertainment                      1.272 0.204911
## studioMiramax                                     1.886 0.060818 .
## studioMiramax Films                               1.984 0.048624 *
## studioMusic Box Films                             2.724 0.007030 **
## studioNelson Entertainment                        0.474 0.636220
## studioNew Line Cinema                             1.372 0.171784
## studioNew Line Home Entertainment                 1.204 0.230003
## studioNew World Pictures                          1.759 0.080205 .
## studioNew Yorker Films                            1.628 0.105221
## studioNewmarket Film Group                        2.209 0.028371 *
## studioNewmarket Films                            -0.535 0.593358
## studioNordisk Film Biograf Distribution           1.254 0.211509
## studioNordisk Film Biografdistributi              2.262 0.024803 *
## studioOctober Films                               1.761 0.079859 .
## studioOpen Road Films                             0.445 0.656893
## studioOrion Home Video                            1.663 0.097955 .
## studioOrion Pictures Corporation                  0.482 0.630383
## studioOutsider Films                              2.590 0.010308 *
## studioOverture Films                              2.429 0.016042 *
## studioParamount                                  -0.977 0.329755
## studioParamount Home Video                        1.799 0.073640 .
## studioParamount Pictures                          1.817 0.070727 .
## studioParamount Studios                           2.260 0.024934 *
## studioRelativity Media                            1.316 0.189811
## studioRoadside Attractions                        2.272 0.024191 *
## studioSag Harbor-Basement Pictures                0.720 0.472367
## studioSaguenay Films                              0.108 0.914314
## studioSamuel Goldwyn Films                        1.807 0.072334 .
## studioScreen Gems                                 0.721 0.471778
## studioSeventh Art Productions                    -0.055 0.956499
## studioSony Pictures                               1.639 0.102746
## studioSony Pictures Classics                      2.199 0.029024 *
## studioSony Pictures Entertainment                 1.400 0.163165
## studioSony Pictures Home Entertainment            1.558 0.120793
## studioSony Pictures/Columbia                      1.493 0.136959
## studioSony Pictures/Screen Gems                   2.451 0.015144 *
## studioStrand Releasing                            0.919 0.359100
## studioSummit Entertainment                        2.955 0.003514 **
## studioTango Entertainment                         0.790 0.430660
## studioThe Film Arcade                             1.624 0.106095
## studioThe Shooting Gallery                        2.316 0.021607 *
## studioThe Weinstein Co.                           2.143 0.033331 *
## studioThe Weinstein Company                       2.624 0.009380 **
## studioThinkFilm                                   1.406 0.161275
## studioTouchstone Home Entertainment               1.726 0.085864 .
## studioTouchstone Pictures                         2.154 0.032440 *
## studioTrimark                                     2.037 0.043024 *
## studioTriStar                                     2.011 0.045723 *
## studioTriStar Pictures                            0.017 0.986137
## studioTwentieth Century Fox Home Entertainment    1.811 0.071708 .
## studioUnited Artists                              3.102 0.002207 **
## studioUniversal Pictures                          2.847 0.004885 **
## studioUniversal Studios                           1.560 0.120318
## studioUrban Vision Entertainment                  3.441 0.000708 ***
```

```
## studioUSA Films                                  1.363 0.174366
## studioVirgin Vision                              1.009 0.314035
## studioWalt Disney Pictures                       1.333 0.184203
## studioWalt Disney Productions                    0.899 0.369854
## studioWarner Bros Pictures                       0.681 0.496609
## studioWarner Bros.                               1.775 0.077429 .
## studioWarner Bros. Pictures                      2.623 0.009407 **
## studioWARNER BROTHERS PICTURES                   0.365 0.715289
## studioWarner Home Video                          1.565 0.119105
## studioWarner Independent                         1.900 0.058912 .
## studioWarner Independent Pictures                1.408 0.160668
## studioWarners Bros. Pictures                     1.528 0.128080
## studioWeinstein Company                          1.645 0.101675
## studioWinstar                                    2.068 0.039916 *
## studioYari Film Group Releasing                  1.970 0.050232 .
## studioZeitgeist Films                            0.451 0.652246
## thtr_rel_year                                   -2.515 0.012708 *
## thtr_rel_month                                   1.082 0.280585
## thtr_rel_day                                     1.385 0.167617
## dvd_rel_year                                    -0.691 0.490386
## dvd_rel_month                                    1.383 0.168224
## dvd_rel_day                                     -0.119 0.905339
## imdb_num_votes                                   6.655 2.80e-10 ***
## best_pic_nomyes                                  2.008 0.046017 *
## best_pic_winyes                                 -3.377 0.000884 ***
## best_actor_winyes                               -0.902 0.368321
## best_actress_winyes                             -1.929 0.055172 .
## best_dir_winyes                                  2.413 0.016764 *
## top200_boxyes                                   -0.305 0.760735
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8005 on 195 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.4843
## F-statistic: 2.968 on 178 and 195 DF,  p-value: 1.464e-13
```

This model has a R-squared value of 0.7304, meaning it explains 73% of the variance.

Implementing AIC, the most significant variables appear to be: title_type, genre, runtime, mpaa_rating, thtr_rel_year, dvd_rel_month, imdb_num_votes, best_pic_nom, best_pic_win and best_dir_win as shown below.

```
set.seed(123)
# use AIC selection model, implementing backward selection
step(lm(imdb_rating~ . , data = training), direction = "backward")
```

```
## Start:  AIC=-51.99
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + studio +
##     thtr_rel_year + thtr_rel_month + thtr_rel_day + dvd_rel_year +
##     dvd_rel_month + dvd_rel_day + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_actor_win + best_actress_win + best_dir_win +
##     top200_box
##
##                     Df Sum of Sq    RSS     AIC
## - studio           147   126.201 251.17 -84.905
## - dvd_rel_day        1     0.009 124.97 -53.961
```

```
## - top200_box         1     0.060 125.02 -53.810
## - dvd_rel_year       1     0.306 125.27 -53.073
## - best_actor_win     1     0.521 125.49 -52.432
## - runtime            1     0.552 125.52 -52.340
## <none>                           124.97 -51.988
## - thtr_rel_month     1     0.750 125.72 -51.749
## - dvd_rel_month      1     1.226 126.19 -50.337
## - thtr_rel_day       1     1.229 126.19 -50.327
## - best_actress_win   1     2.385 127.35 -46.918
## - best_pic_nom       1     2.584 127.55 -46.333
## - mpaa_rating        4     5.203 130.17 -44.732
## - best_dir_win       1     3.730 128.69 -42.987
## - thtr_rel_year      1     4.054 129.02 -42.048
## - title_type         1     4.322 129.29 -41.271
## - best_pic_win       1     7.309 132.27 -32.730
## - genre             10    21.782 146.75 -11.894
## - imdb_num_votes     1    28.379 153.34  22.551
##
## Step:  AIC=-84.9
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     dvd_rel_day + imdb_num_votes + best_pic_nom + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS     AIC
## - dvd_rel_day       1     0.004 251.17 -86.899
## - best_actor_win    1     0.038 251.21 -86.848
## - dvd_rel_year      1     0.055 251.22 -86.823
## - thtr_rel_month    1     0.126 251.29 -86.717
## - best_actress_win  1     0.264 251.43 -86.512
## - top200_box        1     0.691 251.86 -85.878
## - thtr_rel_day      1     0.701 251.87 -85.862
## <none>                          251.17 -84.905
## - best_dir_win      1     1.988 253.15 -83.956
## - runtime           1     2.830 254.00 -82.714
## - best_pic_win      1     3.444 254.61 -81.811
## - title_type        2     4.875 256.04 -81.716
## - dvd_rel_month     1     3.512 254.68 -81.711
## - thtr_rel_year     1     3.883 255.05 -81.167
## - best_pic_nom      1     4.255 255.42 -80.622
## - mpaa_rating       5    10.718 261.88 -79.277
## - genre            10    38.062 289.23 -52.133
## - imdb_num_votes    1    36.135 287.30 -36.632
##
## Step:  AIC=-86.9
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     imdb_num_votes + best_pic_nom + best_pic_win + best_actor_win +
##     best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS     AIC
## - best_actor_win    1     0.039 251.21 -88.841
## - dvd_rel_year      1     0.052 251.22 -88.821
## - thtr_rel_month    1     0.126 251.30 -88.712
```

```
## - best_actress_win  1     0.263 251.43 -88.507
## - top200_box        1     0.695 251.87 -87.866
## - thtr_rel_day      1     0.708 251.88 -87.847
## <none>                          251.17 -86.899
## - best_dir_win      1     1.987 253.16 -85.952
## - runtime           1     2.827 254.00 -84.714
## - best_pic_win      1     3.441 254.61 -83.811
## - title_type        2     4.873 256.04 -83.713
## - dvd_rel_month     1     3.509 254.68 -83.711
## - thtr_rel_year     1     3.881 255.05 -83.164
## - best_pic_nom      1     4.252 255.42 -82.622
## - mpaa_rating       5    10.732 261.90 -81.251
## - genre            10    38.392 289.56 -53.701
## - imdb_num_votes    1    36.136 287.31 -38.627
##
## Step:  AIC=-88.84
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     imdb_num_votes + best_pic_nom + best_pic_win + best_actress_win +
##     best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS     AIC
## - dvd_rel_year      1     0.072 251.28 -90.735
## - thtr_rel_month    1     0.126 251.33 -90.654
## - best_actress_win  1     0.267 251.48 -90.443
## - thtr_rel_day      1     0.699 251.91 -89.802
## - top200_box        1     0.705 251.91 -89.793
## <none>                          251.21 -88.841
## - best_dir_win      1     1.983 253.19 -87.901
## - runtime           1     2.832 254.04 -86.648
## - best_pic_win      1     3.404 254.61 -85.808
## - title_type        2     4.889 256.10 -85.632
## - dvd_rel_month     1     3.658 254.87 -85.434
## - thtr_rel_year     1     4.055 255.26 -84.852
## - best_pic_nom      1     4.218 255.43 -84.613
## - mpaa_rating       5    10.731 261.94 -83.197
## - genre            10    38.358 289.57 -55.696
## - imdb_num_votes    1    36.166 287.38 -40.537
##
## Step:  AIC=-90.73
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_month + imdb_num_votes +
##     best_pic_nom + best_pic_win + best_actress_win + best_dir_win +
##     top200_box
##
##                    Df Sum of Sq    RSS     AIC
## - thtr_rel_month    1     0.109 251.39 -92.573
## - best_actress_win  1     0.268 251.55 -92.336
## - thtr_rel_day      1     0.689 251.97 -91.711
## - top200_box        1     0.763 252.04 -91.602
## <none>                          251.28 -90.735
## - best_dir_win      1     1.947 253.23 -89.847
## - runtime           1     2.912 254.19 -88.426
## - best_pic_win      1     3.336 254.62 -87.803
```

```
## - title_type        2     4.896 256.18 -87.517
## - dvd_rel_month     1     3.606 254.89 -87.405
## - best_pic_nom      1     4.226 255.51 -86.496
## - mpaa_rating       5    10.723 262.00 -85.106
## - thtr_rel_year     1     5.855 257.13 -84.121
## - genre            10    38.444 289.72 -57.492
## - imdb_num_votes    1    36.234 287.51 -42.356
##
## Step:  AIC=-92.57
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_day + dvd_rel_month + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS     AIC
## - best_actress_win  1     0.243 251.63 -94.211
## - top200_box        1     0.759 252.15 -93.445
## - thtr_rel_day      1     0.779 252.17 -93.416
## <none>                          251.39 -92.573
## - best_dir_win      1     1.968 253.36 -91.657
## - runtime           1     3.304 254.69 -89.690
## - title_type        2     4.826 256.21 -89.462
## - best_pic_win      1     3.475 254.87 -89.438
## - dvd_rel_month     1     3.498 254.89 -89.404
## - best_pic_nom      1     4.419 255.81 -88.055
## - mpaa_rating       5    11.223 262.61 -86.237
## - thtr_rel_year     1     5.761 257.15 -86.099
## - genre            10    38.462 289.85 -59.328
## - imdb_num_votes    1    36.416 287.81 -43.977
##
## Step:  AIC=-94.21
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_day + dvd_rel_month + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS     AIC
## - thtr_rel_day     1     0.784 252.42 -95.047
## - top200_box       1     0.792 252.43 -95.035
## <none>                         251.63 -94.211
## - best_dir_win     1     1.889 253.52 -93.413
## - runtime          1     3.124 254.76 -91.596
## - dvd_rel_month    1     3.474 255.11 -91.083
## - title_type       2     4.870 256.50 -91.041
## - best_pic_win     1     3.632 255.26 -90.851
## - best_pic_nom     1     4.347 255.98 -89.806
## - mpaa_rating      5    11.277 262.91 -87.814
## - thtr_rel_year    1     5.762 257.39 -87.744
## - genre           10    38.282 289.92 -61.246
## - imdb_num_votes   1    36.556 288.19 -45.479
##
## Step:  AIC=-95.05
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     dvd_rel_month + imdb_num_votes + best_pic_nom + best_pic_win +
##     best_dir_win + top200_box
##
```

```
##                   Df Sum of Sq    RSS     AIC
## - top200_box       1     0.730 253.15 -95.967
## <none>                         252.42 -95.047
## - best_dir_win     1     1.813 254.23 -94.370
## - runtime          1     3.168 255.58 -92.383
## - dvd_rel_month    1     3.359 255.78 -92.104
## - best_pic_win     1     3.398 255.81 -92.045
## - title_type       2     4.863 257.28 -91.910
## - best_pic_nom     1     4.342 256.76 -90.668
## - mpaa_rating      5    10.688 263.10 -89.538
## - thtr_rel_year    1     5.380 257.80 -89.159
## - genre           10    39.518 291.94 -60.649
## - imdb_num_votes   1    36.502 288.92 -46.533
##
## Step:  AIC=-95.97
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     dvd_rel_month + imdb_num_votes + best_pic_nom + best_pic_win +
##     best_dir_win
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         253.15 -95.967
## - best_dir_win     1     1.801 254.95 -95.316
## - runtime          1     2.987 256.13 -93.580
## - dvd_rel_month    1     3.311 256.46 -93.107
## - title_type       2     4.857 258.00 -92.859
## - best_pic_win     1     3.508 256.65 -92.819
## - best_pic_nom     1     4.599 257.75 -91.234
## - mpaa_rating      5    10.803 263.95 -90.338
## - thtr_rel_year    1     5.249 258.40 -90.291
## - genre           10    39.736 292.88 -61.436
## - imdb_num_votes   1    35.888 289.03 -48.384
##
## Call:
## lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_year + dvd_rel_month + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_dir_win, data = training)
##
## Coefficients:
##                     (Intercept)          title_typeFeature Film
##                       3.225e+01                      -1.021e+00
##             title_typeTV Movie                  genreAnimation
##                      -1.193e+00                       2.490e-01
## genreArt House & International                    genreComedy
##                       1.195e+00                       5.651e-02
##                genreDocumentary                     genreDrama
##                       9.896e-01                       6.822e-01
##                     genreHorror  genreMusical & Performing Arts
##                      -2.751e-01                       1.070e+00
##        genreMystery & Suspense                     genreOther
##                       3.383e-01                       2.756e-01
## genreScience Fiction & Fantasy                        runtime
##                      -6.862e-01                       6.194e-03
##               mpaa_ratingNC-17                   mpaa_ratingPG
```

```
##                     -1.377e-01                     -5.709e-01
##               mpaa_ratingPG-13                   mpaa_ratingR
##                     -8.540e-01                     -4.642e-01
##             mpaa_ratingUnrated                  thtr_rel_year
##                     -2.444e-01                     -1.293e-02
##                  dvd_rel_month                 imdb_num_votes
##                      2.885e-02                      3.721e-06
##                best_pic_nomyes                best_pic_winyes
##                      6.970e-01                     -1.100e+00
##                best_dir_winyes
##                      2.864e-01
```

Next, I create a final model where I linearly regress all the variables displayed above against imdb_rating.

```
set.seed(123)
### final model with what we got implementing AIC
final.model <- lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating +
    thtr_rel_year + dvd_rel_month + imdb_num_votes + best_pic_nom +
    best_pic_win + best_dir_win, data = training)
summary(final.model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_year + dvd_rel_month + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_dir_win, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8713 -0.4072  0.0675  0.5299  2.1397
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.225e+01  9.638e+00   3.346 0.000909 ***
## title_typeFeature Film         -1.021e+00  4.001e-01  -2.551 0.011174 *
## title_typeTV Movie             -1.193e+00  6.414e-01  -1.860 0.063665 .
## genreAnimation                  2.490e-01  4.964e-01   0.502 0.616240
## genreArt House & International  1.195e+00  3.879e-01   3.081 0.002228 **
## genreComedy                     5.651e-02  1.891e-01   0.299 0.765294
## genreDocumentary                9.896e-01  4.338e-01   2.281 0.023145 *
## genreDrama                      6.822e-01  1.629e-01   4.189 3.55e-05 ***
## genreHorror                    -2.751e-01  2.848e-01  -0.966 0.334724
## genreMusical & Performing Arts  1.070e+00  3.703e-01   2.890 0.004094 **
## genreMystery & Suspense         3.383e-01  2.068e-01   1.636 0.102681
## genreOther                      2.756e-01  3.842e-01   0.717 0.473602
## genreScience Fiction & Fantasy -6.862e-01  4.582e-01  -1.498 0.135141
## runtime                         6.194e-03  3.052e-03   2.029 0.043187 *
## mpaa_ratingNC-17               -1.377e-01  9.657e-01  -0.143 0.886681
## mpaa_ratingPG                  -5.709e-01  4.453e-01  -1.282 0.200645
## mpaa_ratingPG-13               -8.540e-01  4.608e-01  -1.853 0.064671 .
## mpaa_ratingR                   -4.642e-01  4.525e-01  -1.026 0.305729
## mpaa_ratingUnrated             -2.444e-01  5.047e-01  -0.484 0.628478
## thtr_rel_year                  -1.293e-02  4.808e-03  -2.690 0.007484 **
## dvd_rel_month                   2.885e-02  1.350e-02   2.136 0.033340 *
## imdb_num_votes                  3.721e-06  5.290e-07   7.034 1.07e-11 ***
```

```
## best_pic_nomyes                6.970e-01  2.768e-01   2.518 0.012252 *
## best_pic_winyes               -1.100e+00  5.000e-01  -2.199 0.028517 *
## best_dir_winyes                2.864e-01  1.818e-01   1.576 0.116041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8517 on 349 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4163
## F-statistic: 12.08 on 24 and 349 DF,  p-value: < 2.2e-16
```

The R-squared value is decreased to 0.45, since we have remove some variables. Even though, the model suggests those eleminated variables were not significant for our prediction, the R-squared value is tended to decrease. Why?

This is normal, since R-square is closely dependent on the number of variables, even though they might not be significant.

Below I tend to choose the best model, using the variables selected by AIC model selection. To compare models, I use ANOVA.

ANOVA

ANOVA is a statistical model which helps us to compare the means of two or more groups. I use ANOVA to analyse the variance in different samples. This is helpful to see how the model changes while I eliminate the least significant variables. During analysing variance, we calculate F statistics, which is similar to other statistics test such as "z" or "t".

```
set.seed(123)
#comparing models using Anova
fit1 <- lm(imdb_rating ~ imdb_num_votes, data = training)
fit2 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes)
fit3 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes + thtr_rel_year)
fit4 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom)
fit5 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
                  best_pic_win + title_type)
fit6 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
                  best_pic_win + title_type + runtime + best_dir_win )
fit7 <- update(modFitAll, imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
                  best_pic_win + title_type  + runtime + best_dir_win + dvd_rel_month + mpaa_rating)
anova(fit1, fit2, fit3, fit4, fit5,fit6, fit7)
```

```
## Analysis of Variance Table
##
## Model 1: imdb_rating ~ imdb_num_votes
## Model 2: imdb_rating ~ genre + imdb_num_votes
## Model 3: imdb_rating ~ genre + imdb_num_votes + thtr_rel_year
## Model 4: imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom
## Model 5: imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
##     best_pic_win + title_type
## Model 6: imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
##     best_pic_win + title_type + runtime + best_dir_win
## Model 7: imdb_rating ~ genre + imdb_num_votes + thtr_rel_year + best_pic_nom +
##     best_pic_win + title_type + runtime + best_dir_win + dvd_rel_month +
##     mpaa_rating
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    372 412.47
```

```
## 2    362 293.72 10   118.749 16.3712 < 2.2e-16 ***
## 3    361 283.58  1    10.144 13.9850 0.0002153 ***
## 4    360 280.13  1     3.454  4.7613 0.0297723 *
## 5    357 272.85  3     7.278  3.3448 0.0193730 *
## 6    355 267.99  2     4.860  3.3499 0.0362190 *
## 7    349 253.15  6    14.841  3.4101 0.0027735 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using ANOVA, I can conclude that all models have a low P-Value. However, I have to choose one of them by checking the normality and comparing the main coefficients.

Checking Normality

Analysis of variance is sensitive to its assumption that model residuals are approximately normal. If they are not, we could get a small p-value for that reason. It is thus worth testing residuals for normality. The Shapiro-Wilk test is quick and easy in R. Normality is its null hypothesis.

```
shapiro.test(fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.95021, p-value = 6.515e-10
```

```
shapiro.test(fit2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.93641, p-value = 1.493e-11
```

```
shapiro.test(fit3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.93192, p-value = 4.902e-12
```

```
shapiro.test(fit4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit4$residuals
## W = 0.92855, p-value = 2.197e-12
```

```
shapiro.test(fit5$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit5$residuals
## W = 0.92678, p-value = 1.455e-12
```

```
shapiro.test(fit6$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit6$residuals
## W = 0.92973, p-value = 2.903e-12
```

```
shapiro.test(fit7$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit7$residuals
## W = 0.93442, p-value = 9.059e-12
```

For all the models, we obtain small p-value (<0.05), so we fail to reject normality. So we can be confident for P-values from F-test results shown in ANOVA test.

Comparing the main coefficients

R-square

R-square values: The last model (fit7) has the highest R-square value since it has more variables than the others. Even though, we keep studying variance, since we can not be confident only by using R-square.

Variance Inflation Factor VIF

A variance inflation factor (VIF) is a ratio of estimated variances, the variance due to including the ith regressor, divided by that due to including a corresponding ideal regressor which is uncorrelated with the others. If a regressor is strongly correlated with others, hence will increase their VIF's, why shouldn't we just exclude it? Excluding it might bias coefficient estimates of regressors with which it is correlated.

```
vif(mod = fit2)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## genre          1.049095 10        1.002399
## imdb_num_votes 1.049095  1        1.024253
```

```
vif(mod = fit3)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## genre          1.131263 10        1.006186
## imdb_num_votes 1.067349  1        1.033126
## thtr_rel_year  1.094167  1        1.046024
```

```
vif(mod = fit4)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## genre          1.163150 10        1.007585
## imdb_num_votes 1.261687  1        1.123248
## thtr_rel_year  1.096779  1        1.047272
## best_pic_nom   1.214672  1        1.102121
```

```
vif(mod = fit5)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## genre          6.308659 10        1.096470
## imdb_num_votes 1.504683  1        1.226655
## thtr_rel_year  1.126928  1        1.061568
## best_pic_nom   1.377962  1        1.173866
## best_pic_win   1.545916  1        1.243349
## title_type     5.601353  2        1.538414
```

```
vif(mod = fit6)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## genre          7.336020 10        1.104773
## imdb_num_votes 1.694898  1        1.301882
## thtr_rel_year  1.186066  1        1.089067
## best_pic_nom   1.414417  1        1.189293
## best_pic_win   1.676705  1        1.294877
## title_type     5.627935  2        1.540236
## runtime        1.568939  1        1.252573
## best_dir_win   1.293505  1        1.137324
```

```
vif(mod = fit7)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## genre          12.911657 10        1.136447
## imdb_num_votes  1.728588  1        1.314758
## thtr_rel_year   1.392840  1        1.180187
## best_pic_nom    1.423430  1        1.193076
## best_pic_win    1.700122  1        1.303887
## title_type      6.268466  2        1.582305
## runtime         1.624759  1        1.274660
## best_dir_win    1.294879  1        1.137928
## dvd_rel_month   1.067919  1        1.033402
## mpaa_rating     3.117694  5        1.120426
```

As expected, as we add more variables, the variance will increase. However, I tend to choose the last model, since it has the highest R-squared value and also the variance does not change drastically compared to the other models. Model 7 (the one suggested by AIC) is the most significant one.

Moreover, I double check the residuals and can see they are all normally distributed and homoskedastic. Check Appendix - Plot 12 to see all the relations Residuals vs Fitted, Normal QQ, Scale-Location and Residuals vs Leverage.

Testing the model

Since I have a small sample, I perform cross Validation:

- Fit the model again on the testing data to obtain the final model.

```
set.seed(123)
mod.testing <-step(lm(imdb_rating~ . , data = testing), direction = "backward")
```

```
## Start:  AIC=-60.57
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + studio +
##     thtr_rel_year + thtr_rel_month + thtr_rel_day + dvd_rel_year +
##     dvd_rel_month + dvd_rel_day + imdb_num_votes + best_pic_nom +
##     best_pic_win + best_actor_win + best_actress_win + best_dir_win +
##     top200_box
```

```
##
##                     Df Sum of Sq     RSS     AIC
## - studio           113    72.858 131.917 -89.676
## - thtr_rel_day       1     0.010  59.069 -62.527
## - thtr_rel_year      1     0.024  59.083 -62.468
## - dvd_rel_day        1     0.032  59.091 -62.436
## - title_type         2     0.526  59.585 -62.394
## - best_actress_win   1     0.119  59.178 -62.075
## - best_pic_nom       1     0.168  59.227 -61.870
## - top200_box         1     0.259  59.318 -61.495
## - thtr_rel_month     1     0.287  59.346 -61.378
## - best_actor_win     1     0.369  59.428 -61.040
## - dvd_rel_month      1     0.400  59.459 -60.912
## <none>                           59.059 -60.567
## - best_dir_win       1     0.645  59.704 -59.905
## - best_pic_win       1     0.703  59.762 -59.668
## - mpaa_rating        4     2.619  61.678 -57.935
## - dvd_rel_year       1     1.235  60.294 -57.495
## - runtime            1     1.297  60.356 -57.246
## - genre             10     7.994  67.053 -49.466
## - imdb_num_votes     1     7.977  67.036 -31.526
##
## Step:  AIC=-89.68
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     dvd_rel_day + imdb_num_votes + best_pic_nom + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win + top200_box
##
##                     Df Sum of Sq    RSS     AIC
## - best_pic_win       1    0.0001 131.92 -91.676
## - runtime            1    0.0224 131.94 -91.635
## - best_actress_win   1    0.0322 131.95 -91.616
## - dvd_rel_day        1    0.0392 131.96 -91.603
## - thtr_rel_month     1    0.0605 131.98 -91.564
## - dvd_rel_month      1    0.0834 132.00 -91.521
## - best_pic_nom       1    0.0876 132.00 -91.514
## - thtr_rel_year      1    0.2189 132.14 -91.270
## - best_actor_win     1    0.2935 132.21 -91.132
## - thtr_rel_day       1    0.3183 132.24 -91.086
## - title_type         2    1.4733 133.39 -90.955
## - best_dir_win       1    0.4978 132.41 -90.753
## - top200_box         1    0.5142 132.43 -90.723
## <none>                          131.92 -89.676
## - mpaa_rating        4    6.4802 138.40 -85.927
## - dvd_rel_year       1    3.9715 135.89 -84.409
## - genre             10   27.9808 159.90 -62.548
## - imdb_num_votes     1   28.5405 160.46 -43.691
##
## Step:  AIC=-91.68
## imdb_rating ~ title_type + genre + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     dvd_rel_day + imdb_num_votes + best_pic_nom + best_actor_win +
##     best_actress_win + best_dir_win + top200_box
##
```

```
##                     Df Sum of Sq    RSS     AIC
## - runtime            1    0.0225 131.94 -93.634
## - best_actress_win   1    0.0322 131.95 -93.616
## - dvd_rel_day        1    0.0393 131.96 -93.603
## - thtr_rel_month     1    0.0604 131.98 -93.564
## - dvd_rel_month      1    0.0836 132.00 -93.521
## - best_pic_nom       1    0.1190 132.04 -93.455
## - thtr_rel_year      1    0.2218 132.14 -93.264
## - best_actor_win     1    0.2936 132.21 -93.131
## - thtr_rel_day       1    0.3200 132.24 -93.082
## - title_type         2    1.4741 133.39 -92.954
## - top200_box         1    0.5141 132.43 -92.723
## - best_dir_win       1    0.5142 132.43 -92.723
## <none>                           131.92 -91.676
## - mpaa_rating        4    6.4811 138.40 -87.926
## - dvd_rel_year       1    4.0662 135.98 -86.238
## - genre             10   27.9855 159.90 -64.540
## - imdb_num_votes     1   28.7970 160.71 -45.300
##
## Step:  AIC=-93.63
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     dvd_rel_day + imdb_num_votes + best_pic_nom + best_actor_win +
##     best_actress_win + best_dir_win + top200_box
##
##                     Df Sum of Sq    RSS     AIC
## - best_actress_win   1    0.0368 131.98 -95.566
## - dvd_rel_day        1    0.0398 131.98 -95.560
## - thtr_rel_month     1    0.0799 132.02 -95.486
## - dvd_rel_month      1    0.0850 132.02 -95.476
## - best_pic_nom       1    0.1200 132.06 -95.411
## - thtr_rel_year      1    0.2402 132.18 -95.189
## - best_actor_win     1    0.3208 132.26 -95.039
## - thtr_rel_day       1    0.3247 132.26 -95.032
## - title_type         2    1.4623 133.40 -94.934
## - top200_box         1    0.5122 132.45 -94.685
## - best_dir_win       1    0.5471 132.49 -94.620
## <none>                           131.94 -93.634
## - mpaa_rating        4    6.4980 138.44 -89.856
## - dvd_rel_year       1    4.1229 136.06 -88.095
## - genre             10   30.2801 162.22 -63.015
## - imdb_num_votes     1   31.1196 163.06 -43.751
##
## Step:  AIC=-95.57
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     dvd_rel_day + imdb_num_votes + best_pic_nom + best_actor_win +
##     best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS     AIC
## - dvd_rel_day      1    0.0396 132.02 -97.492
## - thtr_rel_month   1    0.0682 132.04 -97.439
## - dvd_rel_month    1    0.0799 132.06 -97.418
## - best_pic_nom     1    0.1631 132.14 -97.263
```

```
## - thtr_rel_year   1    0.2275 132.20 -97.144
## - thtr_rel_day    1    0.3069 132.28 -96.997
## - best_actor_win  1    0.3573 132.33 -96.903
## - title_type      2    1.4717 133.45 -96.849
## - best_dir_win    1    0.5248 132.50 -96.593
## - top200_box      1    0.5456 132.52 -96.555
## <none>                        131.98 -95.566
## - mpaa_rating     4    6.4654 138.44 -91.848
## - dvd_rel_year    1    4.3316 136.31 -89.654
## - genre          10   31.0387 163.01 -63.817
## - imdb_num_votes  1   31.1762 163.15 -45.610
##
## Step:  AIC=-97.49
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + thtr_rel_day + dvd_rel_year + dvd_rel_month +
##     imdb_num_votes + best_pic_nom + best_actor_win + best_dir_win +
##     top200_box
##
##                   Df Sum of Sq    RSS     AIC
## - thtr_rel_month  1    0.0723 132.09 -99.358
## - dvd_rel_month   1    0.0813 132.10 -99.341
## - best_pic_nom    1    0.1853 132.20 -99.149
## - thtr_rel_year   1    0.2128 132.23 -99.098
## - thtr_rel_day    1    0.3208 132.34 -98.898
## - best_actor_win  1    0.3535 132.37 -98.837
## - title_type      2    1.4987 133.51 -98.727
## - best_dir_win    1    0.5507 132.57 -98.472
## - top200_box      1    0.5596 132.57 -98.456
## <none>                        132.02 -97.492
## - mpaa_rating     4    6.4721 138.49 -93.766
## - dvd_rel_year    1    4.4729 136.49 -91.329
## - genre          10   31.3126 163.33 -65.346
## - imdb_num_votes  1   31.2620 163.28 -47.422
##
## Step:  AIC=-99.36
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_day + dvd_rel_year + dvd_rel_month + imdb_num_votes +
##     best_pic_nom + best_actor_win + best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - dvd_rel_month   1    0.0581 132.15 -101.251
## - best_pic_nom    1    0.2275 132.32 -100.937
## - thtr_rel_year   1    0.2382 132.33 -100.917
## - thtr_rel_day    1    0.3029 132.39 -100.797
## - best_actor_win  1    0.3762 132.46 -100.662
## - title_type      2    1.6000 133.69 -100.408
## - best_dir_win    1    0.5523 132.64 -100.336
## - top200_box      1    0.6089 132.70 -100.231
## <none>                        132.09  -99.358
## - mpaa_rating     4    6.4082 138.50  -95.752
## - dvd_rel_year    1    4.4210 136.51  -93.292
## - genre          10   31.3051 163.39  -67.249
## - imdb_num_votes  1   31.4130 163.50  -49.087
##
```

```
## Step:  AIC=-101.25
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_day + dvd_rel_year + imdb_num_votes + best_pic_nom +
##     best_actor_win + best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - best_pic_nom     1     0.231 132.38 -102.823
## - thtr_rel_year    1     0.238 132.38 -102.809
## - thtr_rel_day     1     0.303 132.45 -102.689
## - best_actor_win   1     0.367 132.51 -102.570
## - best_dir_win     1     0.519 132.67 -102.290
## - title_type       2     1.641 133.79 -102.226
## - top200_box       1     0.591 132.74 -102.157
## <none>                         132.15 -101.251
## - mpaa_rating      4     6.356 138.50  -97.742
## - dvd_rel_year     1     4.414 136.56  -95.201
## - genre           10    31.344 163.49  -69.104
## - imdb_num_votes   1    31.626 163.77  -50.682
##
## Step:  AIC=-102.82
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_year +
##     thtr_rel_day + dvd_rel_year + imdb_num_votes + best_actor_win +
##     best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - thtr_rel_year    1     0.272 132.65 -104.321
## - thtr_rel_day     1     0.329 132.71 -104.215
## - best_dir_win     1     0.532 132.91 -103.841
## - title_type       2     1.649 134.03 -103.790
## - best_actor_win   1     0.561 132.94 -103.787
## - top200_box       1     0.610 132.99 -103.696
## <none>                         132.38 -102.823
## - mpaa_rating      4     6.324 138.70  -99.389
## - dvd_rel_year     1     4.465 136.84  -96.696
## - genre           10    32.093 164.47  -69.640
## - imdb_num_votes   1    35.007 167.38  -47.337
##
## Step:  AIC=-104.32
## imdb_rating ~ title_type + genre + mpaa_rating + thtr_rel_day +
##     dvd_rel_year + imdb_num_votes + best_actor_win + best_dir_win +
##     top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - thtr_rel_day     1     0.408 133.06 -105.568
## - best_dir_win     1     0.568 133.22 -105.274
## - best_actor_win   1     0.624 133.27 -105.171
## - title_type       2     1.745 134.39 -105.119
## - top200_box       1     0.786 133.43 -104.874
## <none>                         132.65 -104.321
## - mpaa_rating      4     6.507 139.16 -100.588
## - dvd_rel_year     1     9.808 142.46  -88.845
## - genre           10    32.636 165.28  -70.429
## - imdb_num_votes   1    35.653 168.30  -47.997
##
```

```
## Step:  AIC=-105.57
## imdb_rating ~ title_type + genre + mpaa_rating + dvd_rel_year +
##     imdb_num_votes + best_actor_win + best_dir_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - best_dir_win     1     0.598 133.66 -106.469
## - title_type       2     1.709 134.76 -106.442
## - best_actor_win   1     0.632 133.69 -106.408
## - top200_box       1     0.885 133.94 -105.944
## <none>                         133.06 -105.568
## - mpaa_rating      4     6.611 139.67 -101.688
## - dvd_rel_year     1     9.643 142.70  -90.426
## - genre           10    32.441 165.50  -72.113
## - imdb_num_votes   1    35.245 168.30  -49.997
##
## Step:  AIC=-106.47
## imdb_rating ~ title_type + genre + mpaa_rating + dvd_rel_year +
##     imdb_num_votes + best_actor_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - title_type       2     1.708 135.36 -107.357
## - best_actor_win   1     0.747 134.40 -107.103
## - top200_box       1     0.767 134.42 -107.066
## <none>                         133.66 -106.469
## - mpaa_rating      4     6.571 140.23 -102.710
## - dvd_rel_year     1    10.091 143.75  -90.636
## - genre           10    32.178 165.83  -73.618
## - imdb_num_votes   1    35.950 169.60  -50.107
##
## Step:  AIC=-107.36
## imdb_rating ~ genre + mpaa_rating + dvd_rel_year + imdb_num_votes +
##     best_actor_win + top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - best_actor_win   1     0.658 136.02 -108.170
## - top200_box       1     0.776 136.14 -107.956
## <none>                         135.36 -107.357
## - mpaa_rating      4     7.321 142.68 -102.453
## - dvd_rel_year     1    10.233 145.60  -91.503
## - genre           10    44.990 180.35  -57.052
## - imdb_num_votes   1    35.660 171.02  -52.067
##
## Step:  AIC=-108.17
## imdb_rating ~ genre + mpaa_rating + dvd_rel_year + imdb_num_votes +
##     top200_box
##
##                   Df Sum of Sq    RSS      AIC
## - top200_box       1     0.805 136.83 -108.723
## <none>                         136.02 -108.170
## - mpaa_rating      4     7.316 143.34 -103.334
## - dvd_rel_year     1    10.555 146.58  -91.859
## - genre           10    45.517 181.54  -57.449
## - imdb_num_votes   1    35.881 171.90  -52.811
##
```

```
## Step:  AIC=-108.72
## imdb_rating ~ genre + mpaa_rating + dvd_rel_year + imdb_num_votes
##
##                  Df Sum of Sq    RSS      AIC
## <none>                        136.83 -108.723
## - mpaa_rating     4     7.613 144.44 -103.458
## - dvd_rel_year    1    10.550 147.38  -92.526
## - genre          10    45.146 181.97  -58.864
## - imdb_num_votes  1    43.772 180.60  -42.721
```

```
summary(mod.testing)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + mpaa_rating + dvd_rel_year +
##     imdb_num_votes, data = testing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98284 -0.41733  0.04566  0.48191  1.92899
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.062e+02  2.384e+01   4.455 1.31e-05 ***
## genreAnimation                 -7.684e-01  4.576e-01  -1.679 0.094499 .
## genreArt House & International  9.636e-01  3.671e-01   2.625 0.009246 **
## genreComedy                    -1.211e-01  2.080e-01  -0.582 0.560890
## genreDocumentary                1.597e+00  2.996e-01   5.330 2.36e-07 ***
## genreDrama                      7.340e-01  1.767e-01   4.154 4.62e-05 ***
## genreHorror                     2.769e-01  3.109e-01   0.891 0.374068
## genreMusical & Performing Arts  1.550e+00  4.226e-01   3.667 0.000305 ***
## genreMystery & Suspense         6.345e-01  2.426e-01   2.616 0.009500 **
## genreOther                      6.492e-01  3.073e-01   2.113 0.035722 *
## genreScience Fiction & Fantasy  5.534e-01  4.231e-01   1.308 0.192232
## mpaa_ratingPG                  -5.175e-01  2.844e-01  -1.820 0.070087 .
## mpaa_ratingPG-13               -6.487e-01  2.905e-01  -2.233 0.026508 *
## mpaa_ratingR                   -5.035e-01  2.802e-01  -1.797 0.073670 .
## mpaa_ratingUnrated              1.964e-01  3.636e-01   0.540 0.589676
## dvd_rel_year                   -4.991e-02  1.190e-02  -4.193 3.94e-05 ***
## imdb_num_votes                  3.691e-06  4.321e-07   8.540 1.92e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7747 on 228 degrees of freedom
## Multiple R-squared:  0.4526, Adjusted R-squared:  0.4142
## F-statistic: 11.78 on 16 and 228 DF,  p-value: < 2.2e-16
```

The variables with the largest coefficients in the training set, are also displayed through cross validation to be the most significant ones. We have lost some explonatory variables from the training set. So, even though it seems good, it does not solve all the problems. In Appendix - Plot 13, there is a plot showing how the testing data stand when compared to final model. The graph is linearly distributed which shows an acceptable model.

Through AIC, the overfitting is prevented, it is obvious that the value of R square is tended to decrease constantly as we remove variables. Even though our model might be acceptable, there are many trade offs which push me to explore other ways of removing less significant variables. Even though the model studied

so far is not the optimum, it is a convinient and significant model.

Results of Model 1

Below, I display 5 movies and their predicted values. Also, I show the 95% confidence interval values and moreover give the real value of imdb rating. As we can see, the real values tend to be inside the prediction interval and somehow close to predicted value.

```
# Use the final model to generate rating predictions four movies.
set.seed(123)
dataTheDish <- data.frame(title_type ="Feature Film" ,genre="Drama", runtime=101,
                          mpaa_rating="PG-13", dvd_rel_month= 8, best_pic_win="no",
                          best_pic_nom="no", thtr_rel_year=2001,
                          imdb_num_votes=12285, best_dir_win="no")
predTheDish <- predict(final.model, dataTheDish, interval="predict" )

dataFillyBrown <- data.frame(title_type ="Feature Film" ,genre="Drama", runtime=80,
                             mpaa_rating="R", dvd_rel_month= 7, best_pic_win="no",
                             best_pic_nom="no", thtr_rel_year=2001,  imdb_num_votes=899,
                             best_dir_win="no")
predFillyBrown <- predict(final.model, dataFillyBrown, interval="predict")

dataAroundTheWorld <- data.frame(title_type ="Feature Film" ,genre="Action & Adventure",
                                 runtime=120,mpaa_rating="PG",best_pic_win="no",
                                 best_pic_nom="no",thtr_rel_year=2004, dvd_rel_month=11,
                                 imdb_num_votes=66054,best_dir_win="no")
predAroundTheWorld <- predict(final.model, dataAroundTheWorld, interval="predict")

dataHungerGames <- data.frame(title_type ="Feature Film",genre="Drama",runtime=142,
                              mpaa_rating="PG-13",dvd_rel_month= 8, best_pic_win="no",
                              best_pic_nom="no", thtr_rel_year=2012,  imdb_num_votes=675907,
                              best_dir_win="no")
predHungerGames <- predict(final.model, dataHungerGames, interval="predict")

dataTheRaven <- data.frame(title_type ="Feature Film",genre="Mystery & Suspense",runtime=110,
                              mpaa_rating="R",dvd_rel_month= 10, best_pic_win="no",
                              best_pic_nom="no", thtr_rel_year=2012,  imdb_num_votes=71112,
                              best_dir_win="no")
predTheRaven <- predict(final.model, dataTheRaven, interval="predict")
# Show prediction results.
df <- data.frame(t=c("The Dish","Filly Brown","Around the World in 80 Days", "The Hunger Games", "The Ra
                  p=c(sprintf("%2.1f", predTheDish[1]),
                     sprintf("%2.1f", predFillyBrown[1]),
                     sprintf("%2.1f", predAroundTheWorld[1]),
                     sprintf("%2.1f", predHungerGames[1]),
                     sprintf("%2.1f", predTheRaven[1])),
                 i=c(sprintf("%2.1f - %2.1f", predTheDish[2],
                             ifelse(predTheDish[3]>10, 10 ,predTheDish[3] )),
                     sprintf("%2.1f - %2.1f", predFillyBrown[2],
                             ifelse(predFillyBrown[3]>10, 10 ,predFillyBrown[3] )),
                     sprintf("%2.1f - %2.1f", predAroundTheWorld[2],
                             ifelse(predAroundTheWorld[3]>10, 10 ,predAroundTheWorld[3] )),
                     sprintf("%2.1f - %2.1f", predHungerGames[2],
                             ifelse(predHungerGames[3]>10, 10 ,predHungerGames[3] )),
```

```
                    sprintf("%2.1f - %2.1f", predTheRaven[2],
                            ifelse(predTheRaven[3]>10, 10 ,predTheRaven[3] ))),
                  r=c("7.3", "5.5", "5.8", "7.3","6.4"))
kable(df, col.names=c("Movie Title", "Predicted Rating", "95% Prediction Interval", "Actual Rating"))
```

Movie Title	Predicted Rating	95% Prediction Interval	Actual Rating
The Dish	6.1	4.4 - 7.8	7.3
Filly Brown	6.3	4.6 - 8.0	5.5
Around the World in 80 Days	6.0	4.3 - 7.8	5.8
The Hunger Games	8.7	6.9 - 10.0	7.3
The Raven	6.3	4.6 - 8.0	6.4

All the following results can be interpreted as the correlation of each paramenter with the imdb_rating response variable, when all the other attributes remain constant. Most of the attributes were positive correlated with the response variable. All the results below come after studying the final model studied so far using AIC model selection.

(1) The base factor predictor for the genre attribute was Drama and House & International. These genres followed by Musical & Performing Arts tend to have the biggest positive effect on final IMDB rating. Also, documentaries can be significant also according to this model.

(2) The runtime attribute (in hours) got a positive correlation with the imdb_rating.However, we have to be careful with this variable, in the sense that we can not extra extrapolate the fit. Consequently, we can not predict the average imdb_ratings for a 13 hours movie, expecting an incredible high score.

(3) Movies realesed in theaters in a specific time are significant to the model. However, we can not say which years are more successful than the others yet.

(4) One theory that might make sense in a video popularity is that, the more upvotes a movie receives, the more popular it becomes, more people will want to watch it because it is popular and more upvotes it will receive. This makes the variable imdb_num_votes highly affective in the model.

(5) Another important factor is the title type. If a movie belongs to the type Documentary, it is tended to have a higher imdb rating. Then comes Feature Film followed by TV Movie.

(6) Best picture nominees and best director nominees seem to have an effect on the imdb rating also. If a movie is nominated to win a prize, this makes it more enjoyable.

2.2 Model 2 - Feature Modelling and Engineering

Feature engineering is a non standard field of data modelling / machine learning that tries to transform raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Near Zero Variance and Correlated Columns Verification

NZV variables are columns that have little variation among its values, i.e., there is one predominant value, making the values almost constant between the data examples, providing little discriminative power. There are discussions if we should remove the nzv predictors, considering that we are throwing out information but, as we are trying to get the most informative model to our data, we would stick with this approach to remove them.

As for multicollinearity, we will analyze among the predictors if we can find a big correlation between then. Correlated predictors are known to destabilize the model parameters, increase the prediction standard error and worse the fit to new data. Let see how our data is organized. We will use the Caret package to extract the nzv variables and the cor command to extract the correlation matrix:

```
set.seed(123)
nzv <- nearZeroVar(training, saveMetrics = TRUE)
nzvIndices <- which(nzv$nzv == TRUE)
nzv[nzvIndices, ]
```

```
##              freqRatio percentUnique zeroVar  nzv
## best_pic_nom  25.71429     0.5347594   FALSE TRUE
## best_pic_win  73.80000     0.5347594   FALSE TRUE
## top200_box    45.75000     0.5347594   FALSE TRUE
```

As we can see {best_pic_nom, best_pic_win, top200_box} were selected as no discriminative variables. So, the first step is to remove these three variables from further analysis.

Inspecting our correlation matrix, we see that the most correlated columns are between theater_release_year and dvd_rel_year, which makes sense. Anyway, from the correlation model, I choose to keep studying both of them since they correlation is not as strong to be significant. Recheck Appendix - Plot 3. Note that I have removed the critics_score and audience_score from this model, since they are not a subject of my study anymore.

Feature Modeling

In this section, I present the insights that I have made to get to the final model. The movies training dataset was repeatedly sampled and transformed in order to get a high adjusted R^2 score. The list below shows the feature engineering process taken to get to the final model and their respective R^2 scores. I will explain each item and why I have included them in the final model.

- But first, I have to remove the variables analysed above from the training and testing set. After this process, I will have 16 variables to analyse. Also, from the correlation it seems like the theater release day and dvd release day are not so important in the final model. Removing these variables will reduce the number to 14.

```
set.seed(123)
removeVar <- c(8,11, 14,15,19)              # remove thtr_rel_day, dvd_rel_day, best_pic_nom,
                                              # best_pic_win, top200_box

trainingMod <- training[,-removeVar]          # create training set
testingMod <- testing[,-removeVar]            # create testing set
dim(trainingMod); dim(testingMod)
```

```
## [1] 374  14
```

```
## [1] 245  14
```

(1) Transformation of imdb_ratings < 5 to 5: As we are trying to predict the best ratings, I removed the low score ratings and merged then into the 5 score.

(2) Theatre release year transformation: Another feature transformation to try to reduce the variance of the model. Firstly, grouping the years in bins created just 5 years categories, instead of more than 40 years of variations. Secondly, the idea is that maybe the impact of year in the final imdb ratings were not granulated as individual years, but in a range of years.

(3) Inclusion of the biggest studios: Movies coming from the biggests (and more famous) studios could bring some more clues about its ratings. So I selected the most frequent studios in the training dataset

and transformed the other studios to a new category, 'other'. The new added variable was a factor with 6 levels, the 5 biggest studios and the 'other' category.

```
set.seed(123)
moviesVarEngineer <- function(dataFrame, mode){

  # step 1 = Transformation of imdb_ratings < 5 to 5
  dataFrame$imdb_rating[dataFrame$imdb_rating < 5] <- 5

  # step 2 = Theatre release year transformation
  dataFrame$thtr_rel_year_BIN <- cut(dataFrame$thtr_rel_year,
                                     breaks = c(1969, 1980, 1990,2000,2010, 2014))
  dataFrame$thtr_rel_year <- NULL

  # step 3 = selected the  biggest studios and transformed the others to 'other'
  studios <- c('Paramount Pictures','Warner Bros. Pictures','Sony Pictures Home Entertainment',
               'Universal Pictures','Warner Home Video')
  if(mode == 'training'){
    dataFrame$studio <- trainingMod$studio
  }else{
    dataFrame$studio <- testingMod$studio
  }
  dataFrame$studio[!dataFrame$studio %in% studios] <- NA
  dataFrame$studio <- factor(dataFrame$studio, levels = c(studios,'Other'))
  dataFrame$studio[is.na(dataFrame$studio)] <- 'Other'

  invisible(dataFrame)
}
```

Implement the algorithm above into our training data set.

```
set.seed(123)
TrainingPredictors <- moviesVarEngineer(trainingMod,'training')
TestingPredictors <- moviesVarEngineer(testingMod,'testing')

finalModEng <- lm(imdb_rating ~ . , data = TrainingPredictors)
summary(finalModEng)
```

```
##
## Call:
## lm(formula = imdb_rating ~ ., data = TrainingPredictors)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72529 -0.44604 -0.03877  0.48090  1.69473
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          5.853e+01  2.870e+01   2.040
## title_typeFeature Film              -1.059e+00  3.096e-01  -3.420
## title_typeTV Movie                  -8.162e-01  4.944e-01  -1.651
## genreAnimation                       4.606e-02  3.841e-01   0.120
## genreArt House & International       1.023e+00  3.003e-01   3.406
## genreComedy                         -2.197e-04  1.452e-01  -0.002
## genreDocumentary                     7.559e-01  3.349e-01   2.257
```

```
## genreDrama                                5.639e-01  1.270e-01   4.439
## genreHorror                               -2.037e-01  2.202e-01  -0.925
## genreMusical & Performing Arts             8.177e-01  2.856e-01   2.863
## genreMystery & Suspense                    2.134e-01  1.627e-01   1.312
## genreOther                                 3.061e-01  3.019e-01   1.014
## genreScience Fiction & Fantasy            -5.248e-01  3.593e-01  -1.461
## runtime                                    5.879e-03  2.481e-03   2.369
## mpaa_ratingNC-17                          -1.447e-01  7.506e-01  -0.193
## mpaa_ratingPG                             -4.211e-01  3.459e-01  -1.217
## mpaa_ratingPG-13                          -5.585e-01  3.610e-01  -1.547
## mpaa_ratingR                              -2.967e-01  3.530e-01  -0.840
## mpaa_ratingUnrated                        -1.442e-01  3.903e-01  -0.370
## studioWarner Bros. Pictures                4.358e-01  2.088e-01   2.088
## studioSony Pictures Home Entertainment     6.468e-02  2.177e-01   0.297
## studioUniversal Pictures                   4.923e-01  2.130e-01   2.311
## studioWarner Home Video                    2.439e-01  2.426e-01   1.005
## studioOther                                2.239e-01  1.493e-01   1.500
## thtr_rel_month                             1.446e-02  1.057e-02   1.367
## dvd_rel_year                              -2.589e-02  1.435e-02  -1.804
## dvd_rel_month                              1.377e-02  1.074e-02   1.282
## imdb_num_votes                             3.355e-06  3.675e-07   9.131
## best_actor_winyes                         -8.839e-03  1.036e-01  -0.085
## best_actress_winyes                       -1.485e-01  1.157e-01  -1.283
## best_dir_winyes                            2.037e-01  1.347e-01   1.512
## thtr_rel_year_BIN(1980,1990]              -2.325e-01  1.632e-01  -1.425
## thtr_rel_year_BIN(1990,2000]              -5.871e-01  1.485e-01  -3.953
## thtr_rel_year_BIN(2000,2010]              -3.003e-01  1.532e-01  -1.960
## thtr_rel_year_BIN(2010,2014]              -3.334e-01  2.223e-01  -1.500
##                                           Pr(>|t|)
## (Intercept)                               0.042169 *
## title_typeFeature Film                    0.000703 ***
## title_typeTV Movie                        0.099699 .
## genreAnimation                            0.904624
## genreArt House & International            0.000738 ***
## genreComedy                               0.998794
## genreDocumentary                          0.024657 *
## genreDrama                                1.23e-05 ***
## genreHorror                               0.355646
## genreMusical & Performing Arts            0.004460 **
## genreMystery & Suspense                   0.190420
## genreOther                                0.311345
## genreScience Fiction & Fantasy            0.145049
## runtime                                   0.018377 *
## mpaa_ratingNC-17                          0.847218
## mpaa_ratingPG                             0.224293
## mpaa_ratingPG-13                          0.122730
## mpaa_ratingR                              0.401273
## mpaa_ratingUnrated                        0.711984
## studioWarner Bros. Pictures               0.037578 *
## studioSony Pictures Home Entertainment    0.766515
## studioUniversal Pictures                  0.021413 *
## studioWarner Home Video                   0.315440
## studioOther                               0.134446
## thtr_rel_month                            0.172391
```

```
## dvd_rel_year                            0.072160 .
## dvd_rel_month                           0.200783
## imdb_num_votes                           < 2e-16 ***
## best_actor_winyes                       0.932027
## best_actress_winyes                     0.200227
## best_dir_winyes                         0.131442
## thtr_rel_year_BIN(1980,1990]            0.155213
## thtr_rel_year_BIN(1990,2000]            9.38e-05 ***
## thtr_rel_year_BIN(2000,2010]            0.050840 .
## thtr_rel_year_BIN(2010,2014]            0.134538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6495 on 339 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.4998
## F-statistic: 11.96 on 34 and 339 DF,  p-value: < 2.2e-16
```

From this model, the R-square has a value of 0.5454, meaning it explains 54% of the variance. This is improved compared to the previous model, when we eleminated variables using AIC variable selection model. However, we can not say that we have reached the optimum yet. There will always be trade offs among these models.

The model diagnostic plots in Appendix - Plot 14 show that the model is acceptable. There is good scatter of the residuals around zero for the range of fitted values (the mean value of the residuals is, in fact, zero). The residuals distribution histogram show a pretty normal distribution.

Overall, the evidence points toward the final model being valid.

Testing and Evaluation

Declaring the model efficiency by the estimated training error could be a little unrealistic because the model tries to minimize the training set error. A good way to evaluate this regression model is the Root Mean Square Error - RMSE, which makes an estimate of the standard error made by the model in unseen data.

```
set.seed(123)
testingPredictions  <- predict(finalModEng, TestingPredictors)
trainingPredictions  <- predict(finalModEng, TrainingPredictors)

postResample(pred=testingPredictions ,obs=TestingPredictors$imdb_rating)[1]
```

```
##      RMSE
## 0.6694469
```

```
postResample(pred=trainingPredictions ,obs=TrainingPredictors$imdb_rating)[1]
```

```
##     RMSE
## 0.618349
```

The feature engineering process was the responsible in greatly improving the adjusted R^2 and the model seemed to generalize well in new unseen data, as we obtained a relatively similar RMSE for training and testing set. This makes our model acceptable.

Moreover in Appendix - Plot 15, there is another plot showing how the testing data stand when compared to final model engineering. The graph is linearly distributed which shows an acceptable model.

Results on Model 2

Below, I display the same 5 movies showed above and their predicted values. Also, I show the 95% confidence interval values and moreover give the real value of imdb rating.

```
set.seed(123)
# Use the final model to generate rating predictions four movies.
dataTheDish <- data.frame(title_type ="Feature Film" ,genre="Drama", runtime=101,
                          mpaa_rating="PG-13",studio= "Warner Bros. Pictures",
                          thtr_rel_month=3, dvd_rel_month= 8, best_actor_win="no",
                          best_actress_win="no", thtr_rel_year_BIN='(2000,2010]',
                          imdb_num_votes=12285,
                          best_dir_win="no",dvd_rel_year=2013)
predTheDish <- predict(finalModEng, dataTheDish, interval="predict")

dataFillyBrown <- data.frame(title_type ="Feature Film" ,genre="Drama", runtime=80,
                             mpaa_rating="R",studio= "Other", thtr_rel_month=4,
                             dvd_rel_month= 7, best_actor_win="no", best_actress_win="no",
                             thtr_rel_year_BIN='(2010,2014]',
                             imdb_num_votes=899,best_dir_win="no",
                             dvd_rel_year=2001)
predFillyBrown <- predict(finalModEng, dataFillyBrown, interval="predict")

dataAroundTheWorld <- data.frame(title_type ="Feature Film" ,genre="Action & Adventure",
                                 runtime=120, mpaa_rating="PG",studio= "Other",
                                 thtr_rel_month=6, dvd_rel_month= 11,
                                 best_actor_win="no", best_actress_win="no",
                                 thtr_rel_year_BIN='(2000,2010]',
                                 imdb_num_votes=66054,best_dir_win="no",
                                 dvd_rel_year=2004)
predAroundTheWorld <- predict(finalModEng, dataAroundTheWorld, interval="predict")

dataHungerGames <- data.frame(title_type ="Feature Film" ,genre="Drama", runtime=142,
                              mpaa_rating="PG-13",studio= "Other",
                              thtr_rel_month=3,
                              dvd_rel_month= 8, best_actor_win="no",
                              best_actress_win="yes",
                              thtr_rel_year_BIN='(2010,2014]',  imdb_num_votes=675907,
                              best_dir_win="no", dvd_rel_year=2012)
predHungerGames <- predict(finalModEng, dataHungerGames, interval="predict")

dataTheRaven <- data.frame(title_type ="Feature Film" ,genre="Mystery & Suspense", runtime=110,
                              mpaa_rating="R",studio= "Other",
                              thtr_rel_month=4,
                              dvd_rel_month= 10, best_actor_win="no",
                              best_actress_win="no",
                              thtr_rel_year_BIN='(2010,2014]',  imdb_num_votes=71112,
                              best_dir_win="no", dvd_rel_year=2012)
predTheRaven <- predict(finalModEng, dataTheRaven, interval="predict")

# Show prediction results.
df <- data.frame(t=c("The Dish","Filly Brown","Around the World in 80 Days", "The Hunger Games", "The Ra
                 p=c(sprintf("%2.1f", predTheDish[1]),
                     sprintf("%2.1f", predFillyBrown[1]),
```

```
                    sprintf("%2.1f", predAroundTheWorld[1]),
                    sprintf("%2.1f", predHungerGames[1]),
                    sprintf("%2.1f", predTheRaven[1])),
                i=c(sprintf("%2.1f - %2.1f", predTheDish[2], predTheDish[3]),
                    sprintf("%2.1f - %2.1f", predFillyBrown[2], predFillyBrown[3]),
                    sprintf("%2.1f - %2.1f", predAroundTheWorld[2], predAroundTheWorld[3]),
                    sprintf("%2.1f - %2.1f", predHungerGames[2], predHungerGames[3]),
                    sprintf("%2.1f - %2.1f", predTheRaven[2], predTheRaven[3])),
                r=c("7.3", "5.5", "5.8", "7.3", "6.4"))
kable(df, col.names=c("Movie Title", "Predicted Rating", "95% Prediction Interval", "Actual Rating"))
```

Movie Title	Predicted Rating	95% Prediction Interval	Actual Rating
The Dish	6.3	4.9 - 7.6	7.3
Filly Brown	6.5	5.1 - 7.8	5.5
Around the World in 80 Days	6.3	4.9 - 7.6	5.8
The Hunger Games	8.4	7.0 - 9.8	7.3
The Raven	6.3	5.0 - 7.6	6.4

All the following results can be interpreted as the correlation of each paramenter with the imdb_rating response variable, when all the other attributes remain constant. Most of the attributes were positive correlated with the response variable. All the results below come after studying the Model Engineering.

(1) The base factor predictor for the genre attribute was Drama and House & International. These genres followed by Musical & Performing Arts tend to have the biggest positive effect on final IMDB rating. Both models studied, output the same results on genres and their effectivness on IMDB rating.

(2) The runtime attribute (in hours) got a positive correlation with the imdb_rating.However, we have to be careful with this variable, in the sense that we can not extra extrapolate the fit. Consequently, we can not predict the average imdb_ratings for a 13 hours movie, expecting an incredible high score. The same result is displayed in the previous model also. But in this case, runtime has a more significant coefficient.

(3) Movies realesed in the last decade of the last century tend to be more successful than the others. If the theater release year is between 1990 and 2000, this makes a movie enjoyable. In the previous model, we can not make such a detailed analysis in the theater released year.

(4) One theory that might make sense in a video popularity is that, the more upvotes a movie receives, the more popular it becomes, more people will want to watch it because it is popular and more upvotes it will receive. This makes the variable imdb_num_votes highly affective in the model. This results correlate to the previous model also.

(5) Another important factor is the title type. If a movie belongs to the type Feature Film, it is tended to have a higher imdb rating. This result differs from the previous model, where documentary type was the most influential variable.

(6) Studio of production is very important in this model. Movies produced in 'Bros. Pictures' and 'Universal Pictures' seem to have a effect on the final IMDB rating. In the previous model, Studio was eleminated from studying as insignificant.

3. Conclusion

On this project, I tried to make a deep analysis and prediction on movies. I started by implementing some extended exploratory data analysis, in order to understand the dataset better. Mostly, I studied the correlation among variables and chose imdb_rating to be my dependent variable.

Next, I splitted the dataset into a training and testing set, so I could have the chance to test my prediction models. After this, I tried to build two prediction models. The first one was using the standard linear model function and the stepwise automated model selection algorithm, more specifically the AIC stepwise model selection. After this process, I constructed some accuracy and normality testing. In overall, the process was acceptable, and the variables selected were significant to the model.

The second model was built based on the feature engineering process. The feature engineering process was responsible in improving the adjusted R2 and the model seemed to generalize well in new unseen data, as we obtained a relatively low RMSE. I implemented some feature transformation, so the variables could have been more significant. The NZV and Correlation analysis taken with the final model eleminated some probable dependent columns. After selecting the final model, I tested the model, by comparing the RMSE for both testing and training set. They had almost similar values, so I can say the model responded positively.

In overall, both models were successful to make predictions. The most influential variables in the first model were significant in the second one also. If I have to pick among them, I would say Model Engineering. Feature engineering tries to transform raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Before the first raw model, I managed to extract a better model with the stepwise selection, after that, the criterious feature selection, transformation and engineering processes always gave a better fit in the training set.

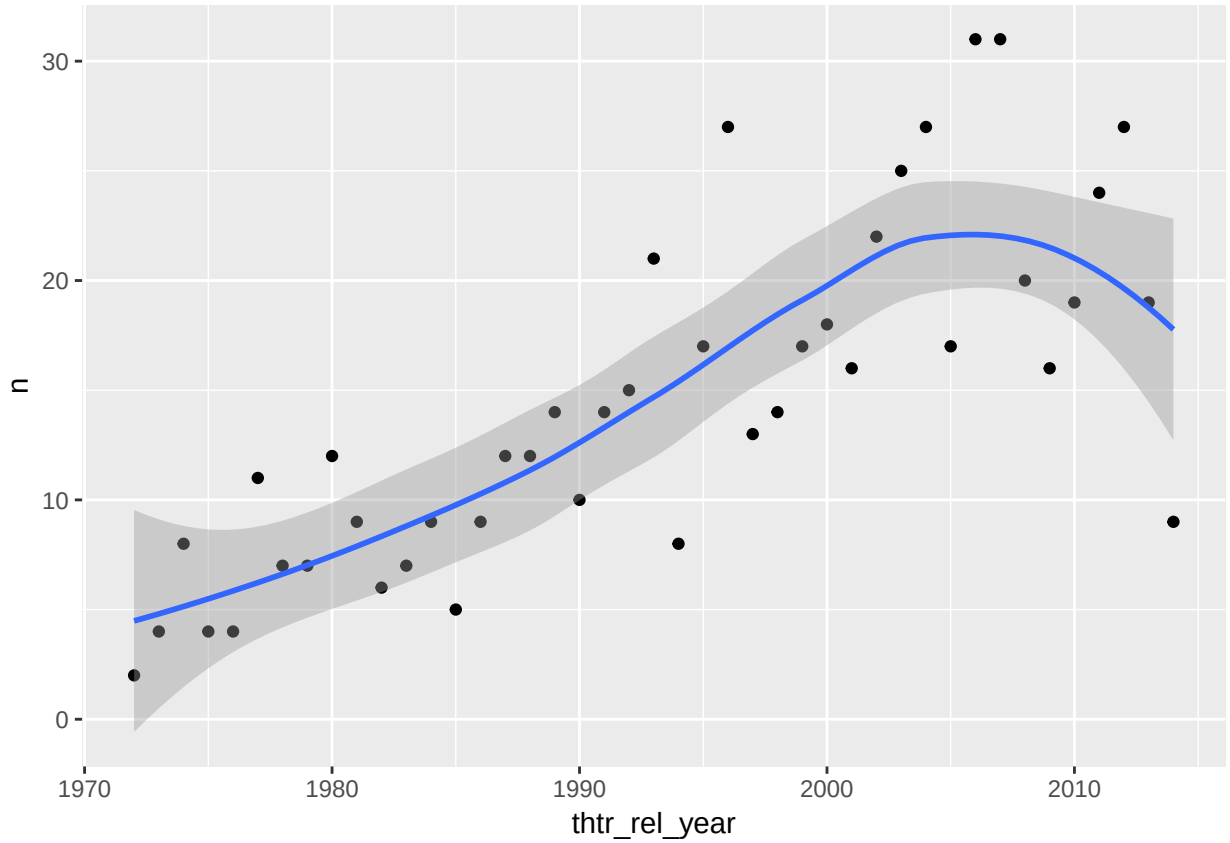
In this project I picked 'imdb_rating' variable as a response variable. Similar analysis could be made using the other two continuous variables ('critics_score' and 'audience_score'). Since all these three variables were highly correlated, I expect a similar response even if you consider critics_score or audience_score as a dependent variable. Moreover, I would like to mention that I tried to use support vector machines predicting the models, instead of linear models. I chose to work on linear models, because the model accuracy was almost the same in both cases. Linear Models are the most popular method on predicting multivariate covariates, but I wouldn't exlude any other option. Since the results were almost the same, I prefered to use linear regression models in this case.

There is a lot of research that can be done and the one explained in this project shows that we can accurately predict how people will rate the movies from its general characteristics.

Appendix

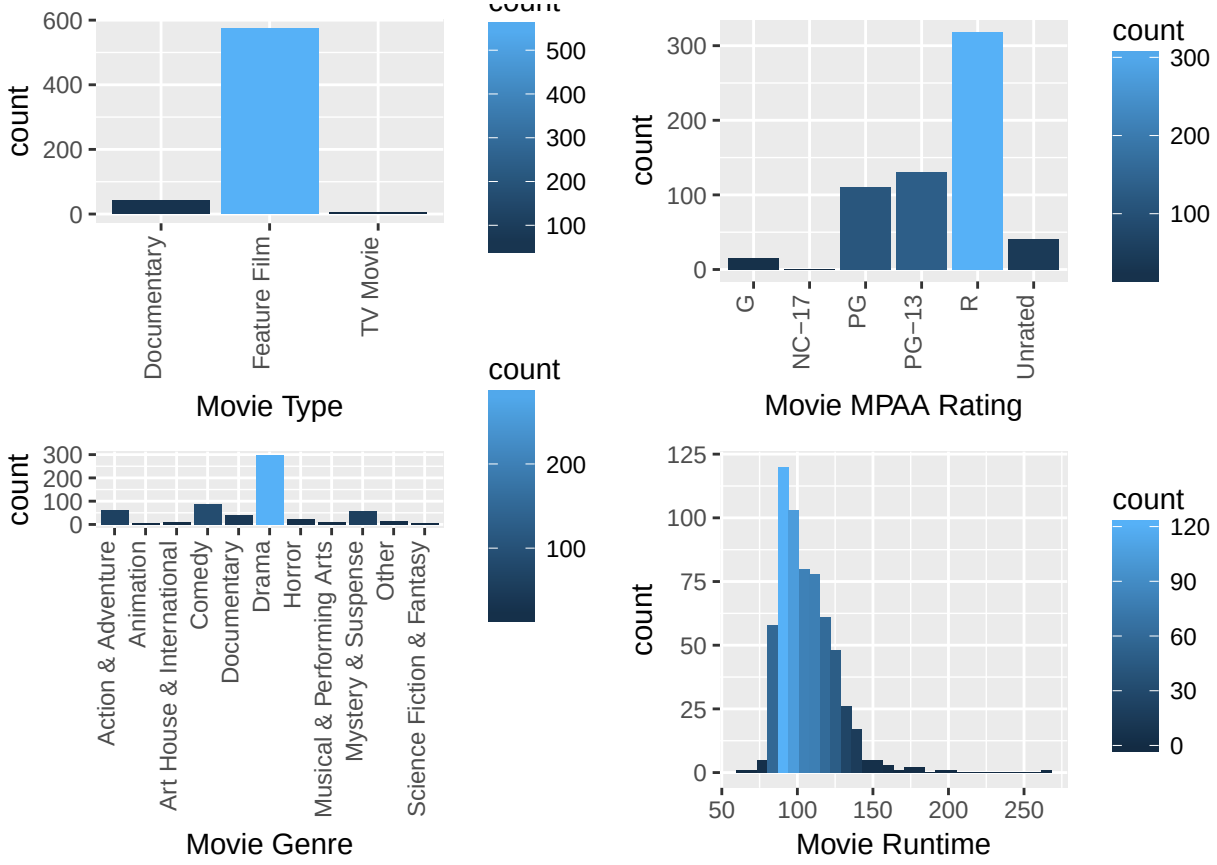
Plot 1 - Movies Trend per year

```
# ggplot showing the trend of movies per year
temp <- movies %>% select(title,thtr_rel_year) # select only movie title and year released
temp <- temp %>% group_by(thtr_rel_year) %>% summarise(n=n()) # how many movies produced in a specific
temp <- na.omit(temp)
ggplot(temp, aes(thtr_rel_year, n)) +  geom_point() +  geom_smooth()
```



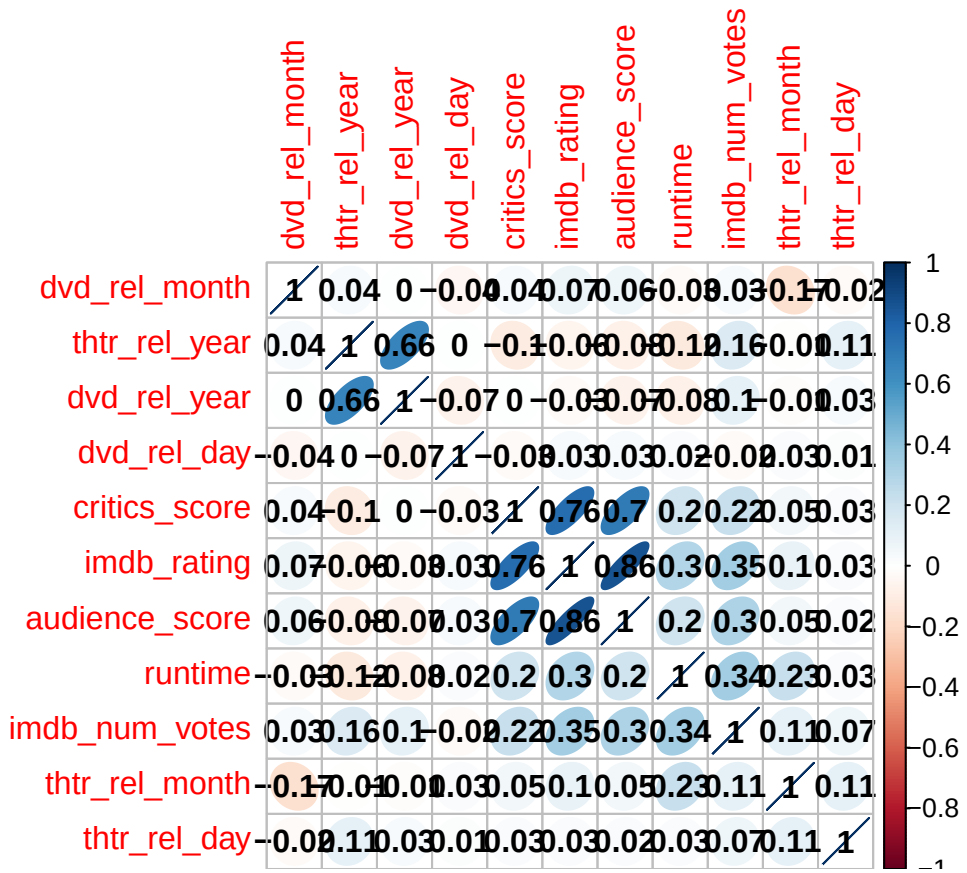
Plot 2 - Charts of data distribution

```
# Create histograms of some of the key movie characteristic data.
p1 <- ggplot(data=movies, aes(x=genre)) +
      geom_bar(aes(fill = ..count..)) +
      xlab("Movie Genre") +
      theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p2 <- ggplot(data=movies, aes(x=title_type)) +
      geom_bar(aes(fill = ..count..)) +
      xlab("Movie Type") +
      theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p3 <- ggplot(data=movies, aes(x=mpaa_rating)) +
      geom_bar(aes(fill = ..count..)) +
      xlab("Movie MPAA Rating") +
      theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))
p4 <- ggplot(data=movies, aes(x=runtime)) +
      geom_histogram( aes(fill = ..count..)) +
      xlab("Movie Runtime")
grid.arrange(p2, p3, p1, p4, nrow=2)
```



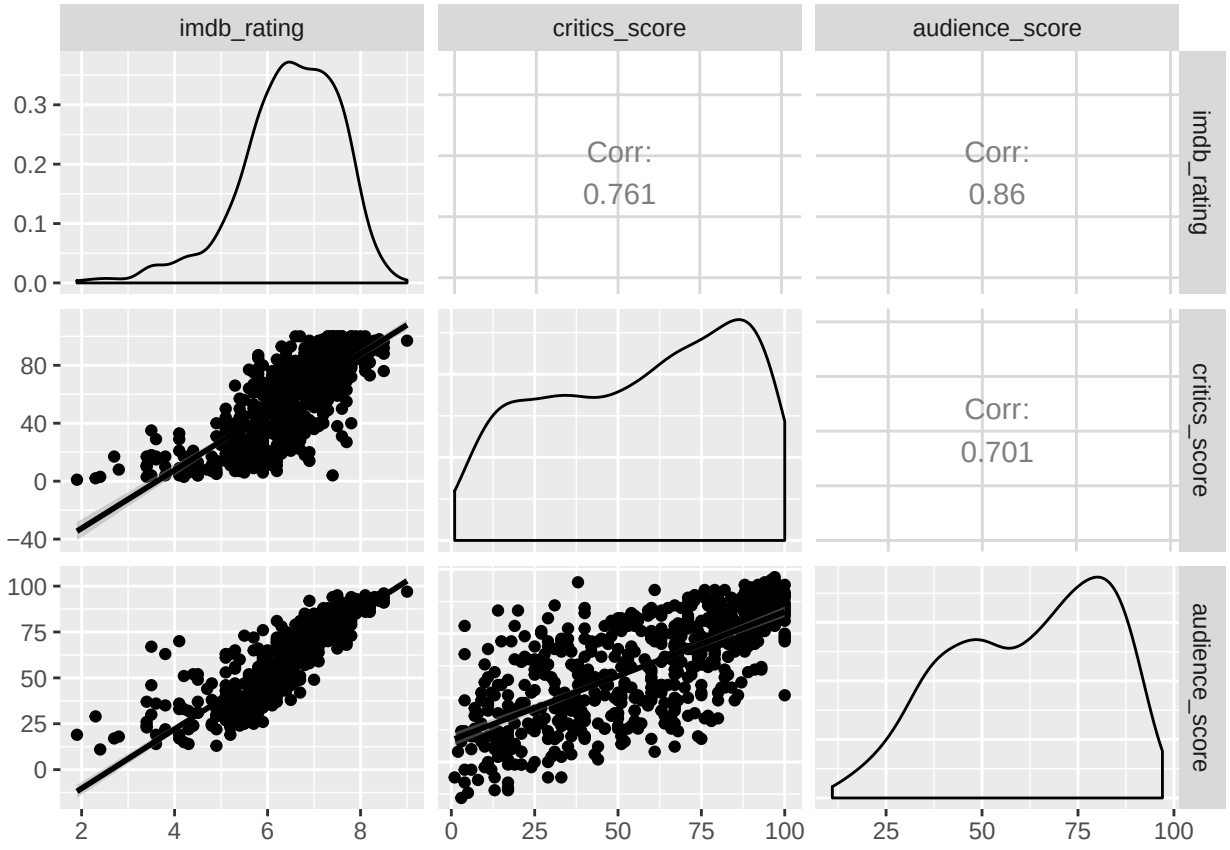
Plot 3 - Correlation of all numeric variables.

```
# Data cleaning
movies <- na.omit(movies)
movies <- unique(movies)
# Correlation between numeric values
correlation <- cor(movies[sapply(movies, is.numeric)])
corrplot(correlation, order = "hclust", method = "ellipse", addCoef.col = "black")
```



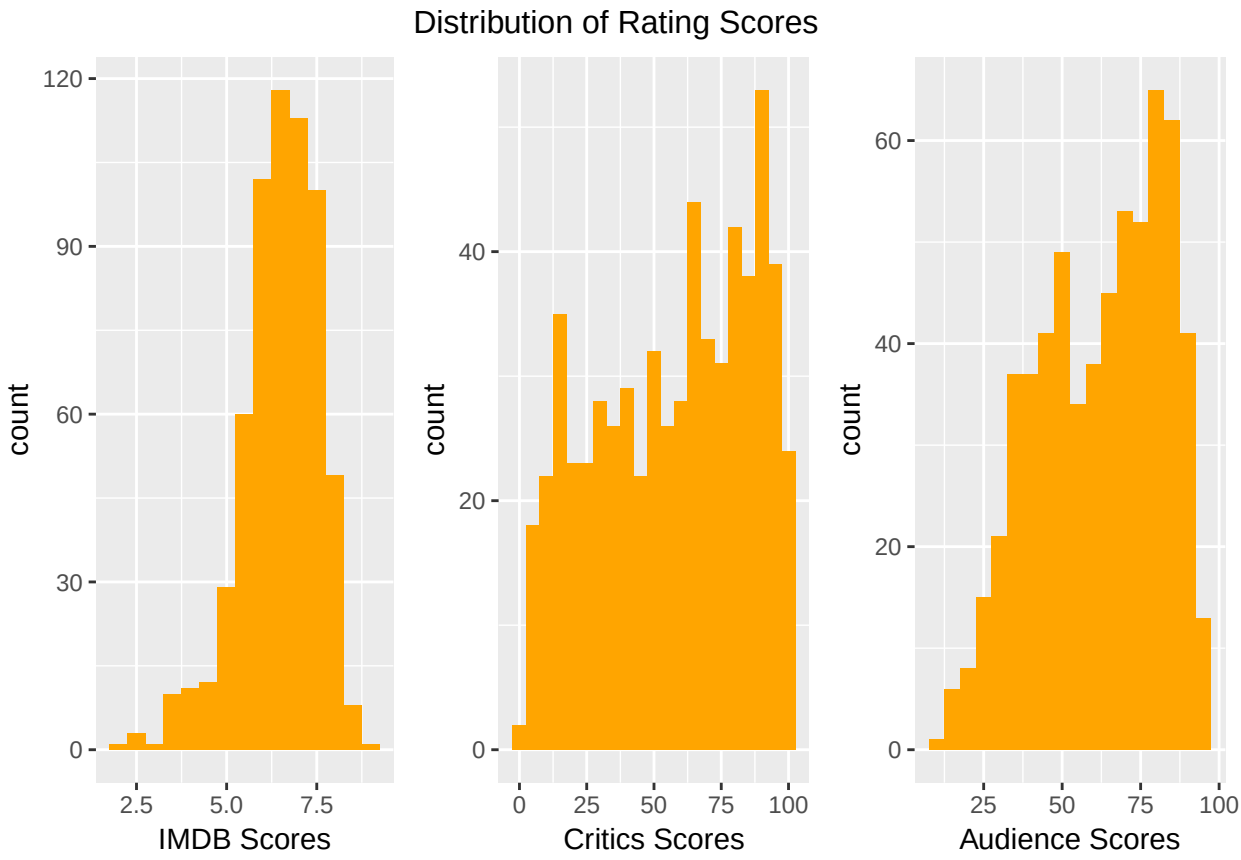
Plot 4 - Pairwise plots of movie rating scores

```
# Create pairwise plots of the movie rating scores to test for colinearity.
responseCols <- c(13,16,18)
movies1 <- movies[,responseCols]
g = ggpairs(movies1, lower = list(continuous = "smooth"),method = "loess")
g
```



Plot 5 - Histograms of movie rating scores

```
# Create histograms of the movie rating scores.
p1 <- ggplot(data=movies, aes(x=imdb_rating)) +
      geom_histogram(binwidth=0.5, fill="orange") +
      xlab("IMDB Scores")
p2 <- ggplot(data=movies, aes(x=critics_score)) +
      geom_histogram(binwidth=5, fill="orange") +
      xlab("Critics Scores")
p3 <- ggplot(data=movies, aes(x=audience_score)) +
      geom_histogram(binwidth=5, fill="orange") +
      xlab("Audience Scores")
grid.arrange(p1, p2, p3, nrow=1,
             top="Distribution of Rating Scores")
```



Plot 6 - Correlation of IMDB rating to other variables

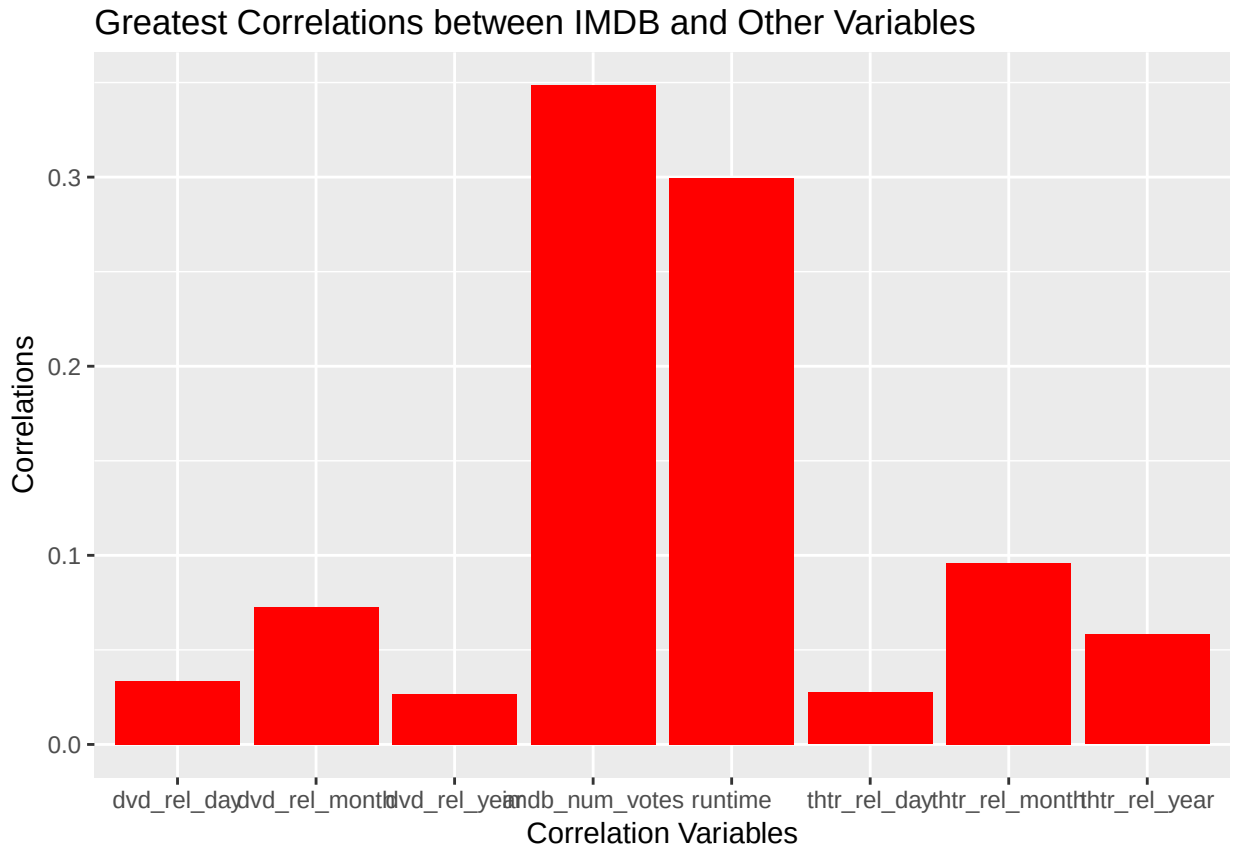
```
#remove critics_score and audience_score variables since we have studied them
movies$critics_score <- NULL
movies$audience_score <- NULL

#select numeric columns
movies %>% remove_missing()
```

```
## # A tibble: 618 × 30
##                      title   title_type              genre runtime
##                      <chr>       <fctr>             <fctr>   <dbl>
## 1              Filly Brown Feature Film              Drama      80
## 2                 The Dish Feature Film              Drama     101
## 3     Waiting for Guffman Feature Film             Comedy      84
## 4    The Age of Innocence Feature Film              Drama     139
## 5              Malevolence Feature Film             Horror      90
## 6                Lady Jane Feature Film              Drama     142
## 7             Mad Dog Time Feature Film              Drama      93
## 8  Beauty Is Embarrassing  Documentary        Documentary      88
## 9    The Snowtown Murders Feature Film              Drama     119
## 10            Superman II Feature Film Action & Adventure     127
## # ... with 608 more rows, and 26 more variables: mpaa_rating <fctr>,
## #   studio <fctr>, thtr_rel_year <dbl>, thtr_rel_month <dbl>,
## #   thtr_rel_day <dbl>, dvd_rel_year <dbl>, dvd_rel_month <dbl>,
```

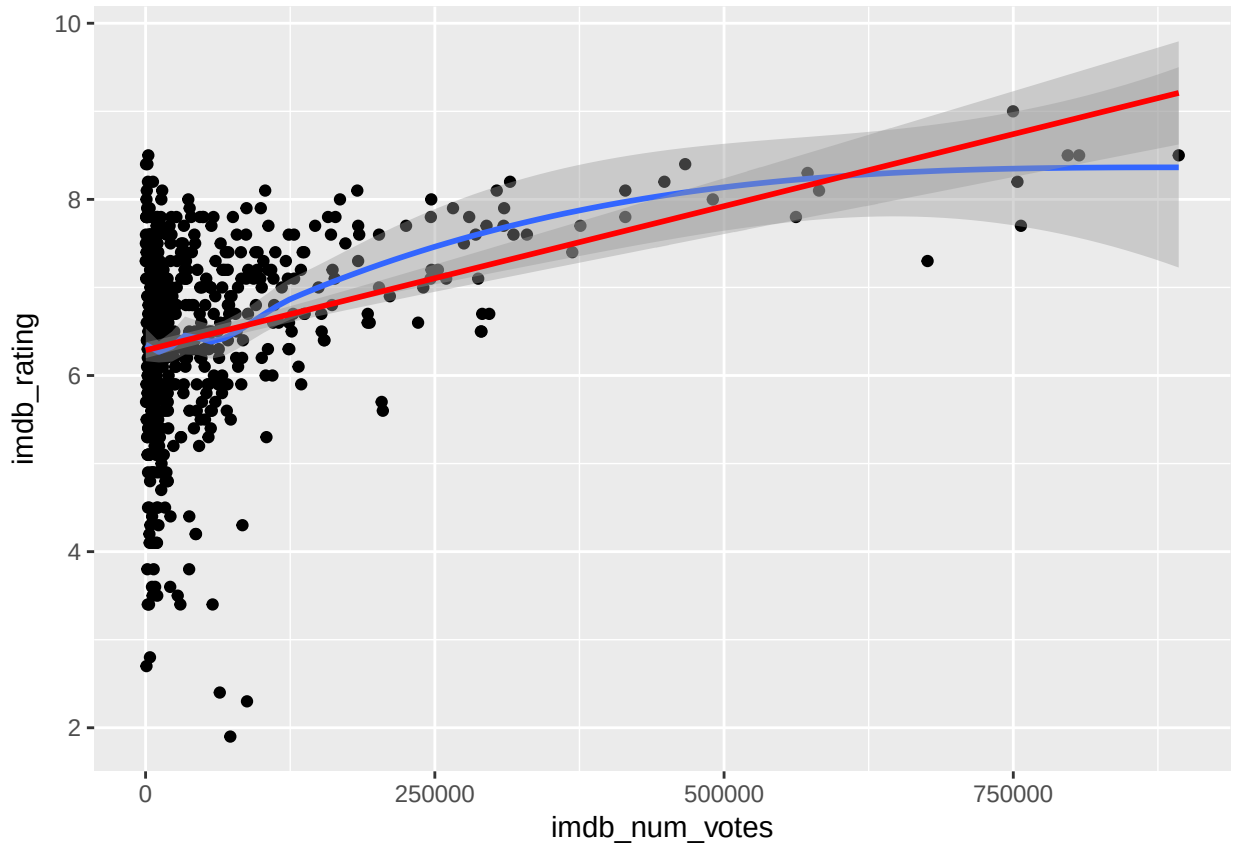
```
## #   dvd_rel_day <dbl>, imdb_rating <dbl>, imdb_num_votes <int>,
## #   critics_rating <fctr>, audience_rating <fctr>, best_pic_nom <fctr>,
## #   best_pic_win <fctr>, best_actor_win <fctr>, best_actress_win <fctr>,
## #   best_dir_win <fctr>, top200_box <fctr>, director <chr>, actor1 <chr>,
## #   actor2 <chr>, actor3 <chr>, actor4 <chr>, actor5 <chr>,
## #   imdb_url <chr>, rt_url <chr>
```

```
nums <- sapply(movies, is.numeric)
#and here you have "the most important correlations" for variable cases excluding character variables
Movie_Corr <- movies[,nums]  %>% correlate() %>% focus(imdb_rating)
ggplot(data=Movie_Corr, aes(x=rowname, y=abs(imdb_rating))) +
  geom_bar(stat="identity", position="identity", fill = "red") +
    scale_colour_solarized("red") +
  ylab("Correlations") +
  xlab("Correlation Variables") +
  ggtitle("Greatest Correlations between IMDB and Other Variables")
```



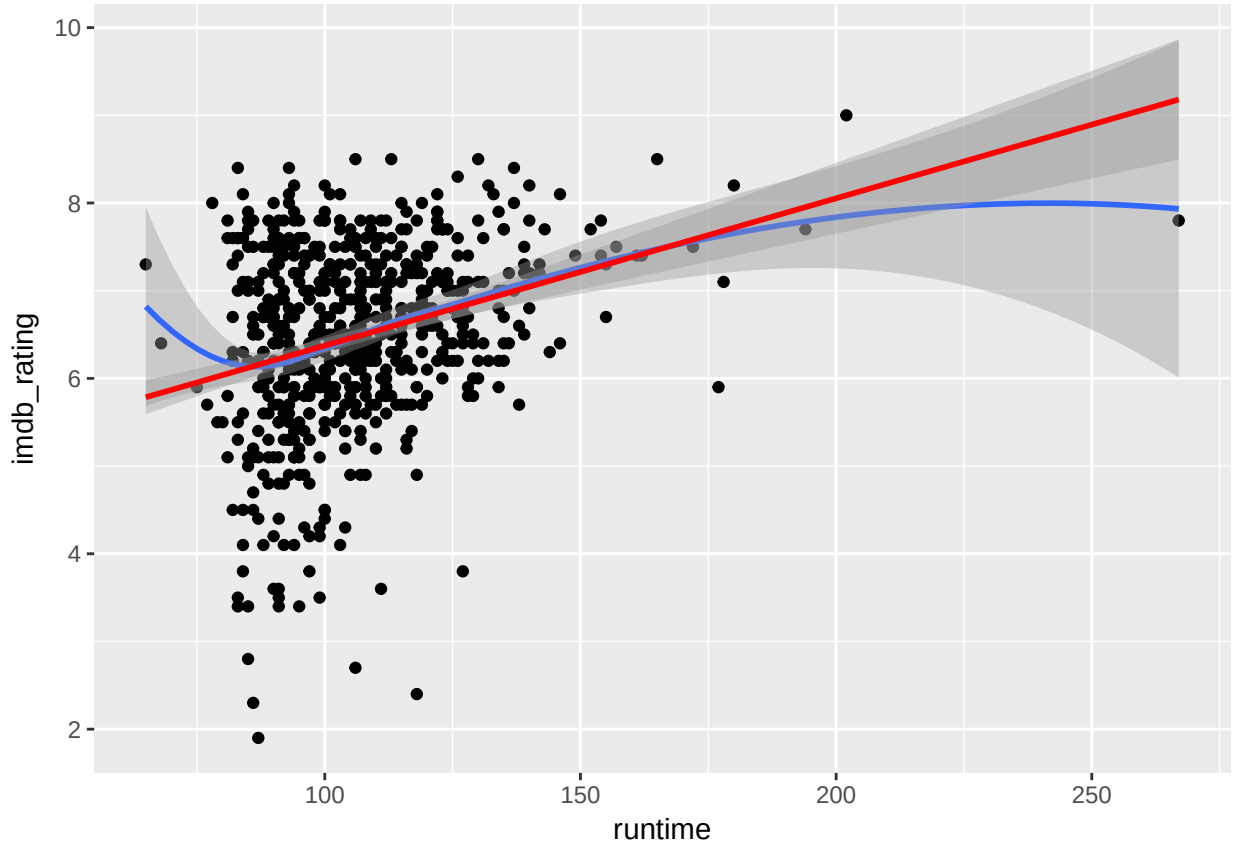
Plot 7 - Number of users versus IMDB score

```
ggplot(movies, aes(x = imdb_num_votes, y = imdb_rating)) + geom_point() + geom_smooth() +
    geom_smooth(method= "lm", color = "red")
```



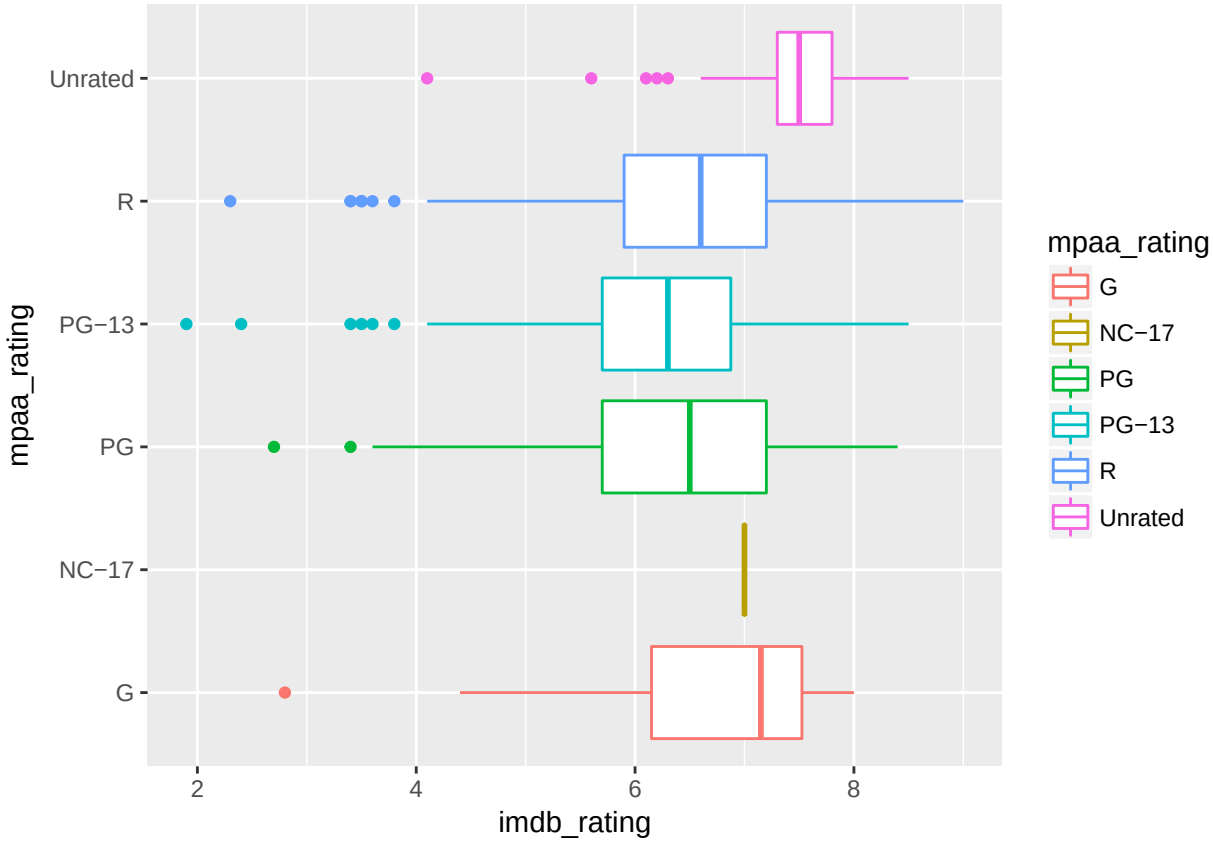
Plot 8 - Runtime versus IMDB rating

```
ggplot(movies, aes(x = runtime, y = imdb_rating)) + geom_point() + geom_smooth() +
    geom_smooth(method= "lm", color = "red")
```



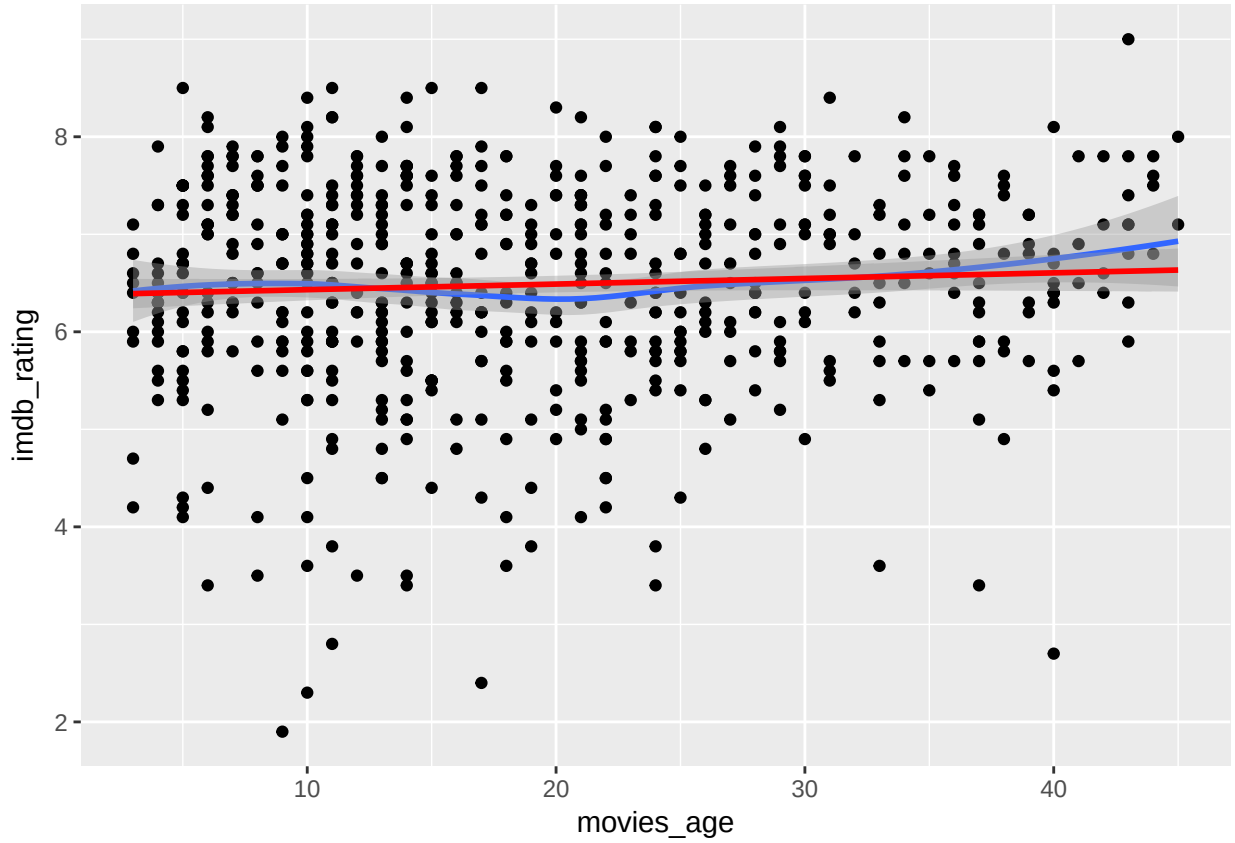
Plot 9 - IMDB rating versus MPAA rating

```
# plot imdb rating based on mpaa_rating
rating <- movies %>% select(imdb_rating,mpaa_rating)
rating <- na.omit(rating)
rate <- ggplot(rating, aes(x=mpaa_rating, y = imdb_rating, colour=mpaa_rating)) +
    geom_boxplot() + coord_flip()
rate
```



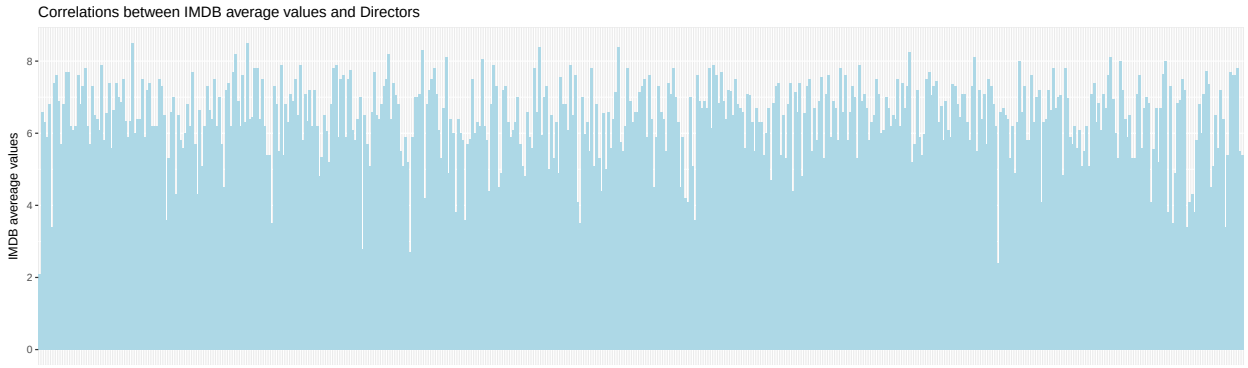
Plot 10 - Movies age vesus IMDB rating

```
# find movies age
current_year <- format(Sys.time(), " %Y")
movies$movies_age <- as.numeric(current_year)- movies$thtr_rel_year
# plot the relation
Score_Age <- ggplot(movies, aes(x = movies_age, y = imdb_rating)) + geom_point() +
    geom_smooth() +geom_smooth(method= "lm", color = "red")
Score_Age
```



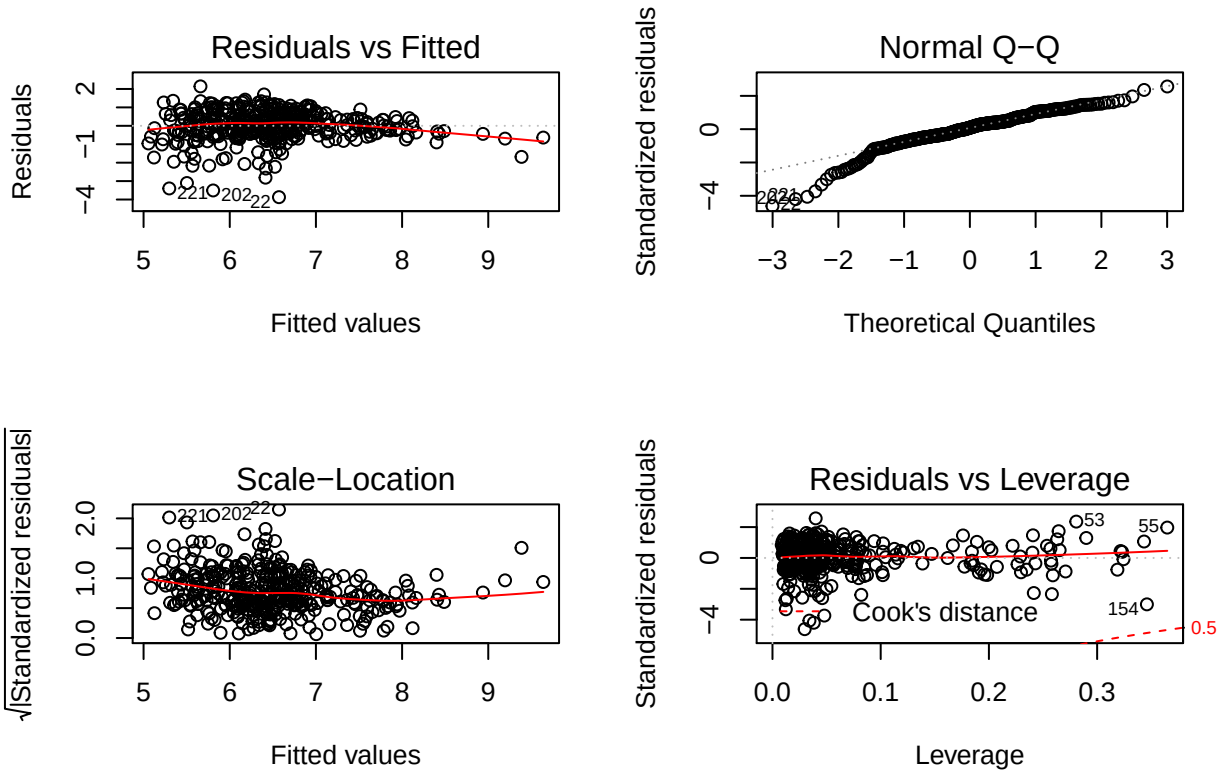
Plot 11 - Directors versus IMDB rating

```
## relationship of average imdb rating for each director
director <- movies %>% select(director,imdb_rating)
director <- director %>% group_by(director) %>% summarise(avg_imdb_rating=mean(imdb_rating))
director <- director %>% arrange(desc(avg_imdb_rating))
ggplot(data=director, aes(x=director, y=avg_imdb_rating)) +
  geom_bar(stat="identity", position="identity", fill = "lightblue") +
    scale_colour_solarized("blue") +
  ylab("IMDB avereage values") +
   ggtitle("Correlations between IMDB average values and Directors") +
    theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Plot 12 - Residuals and Diagnostics for Model 1

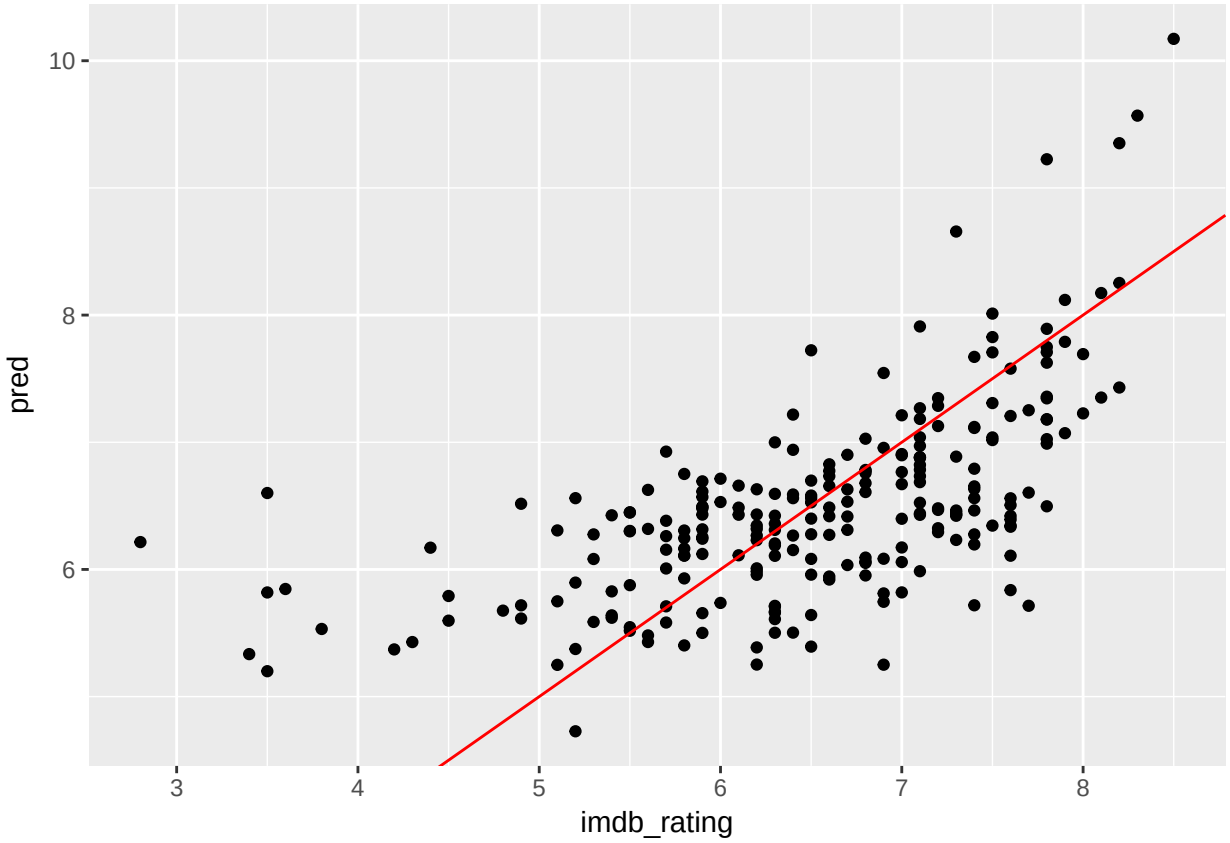
```
# check residuals and diagnostics for the final model in AIC
par(mfrow = c(2, 2))
plot(final.model)
```



Plot 13 - Predicted versus truth in test set for Model 1

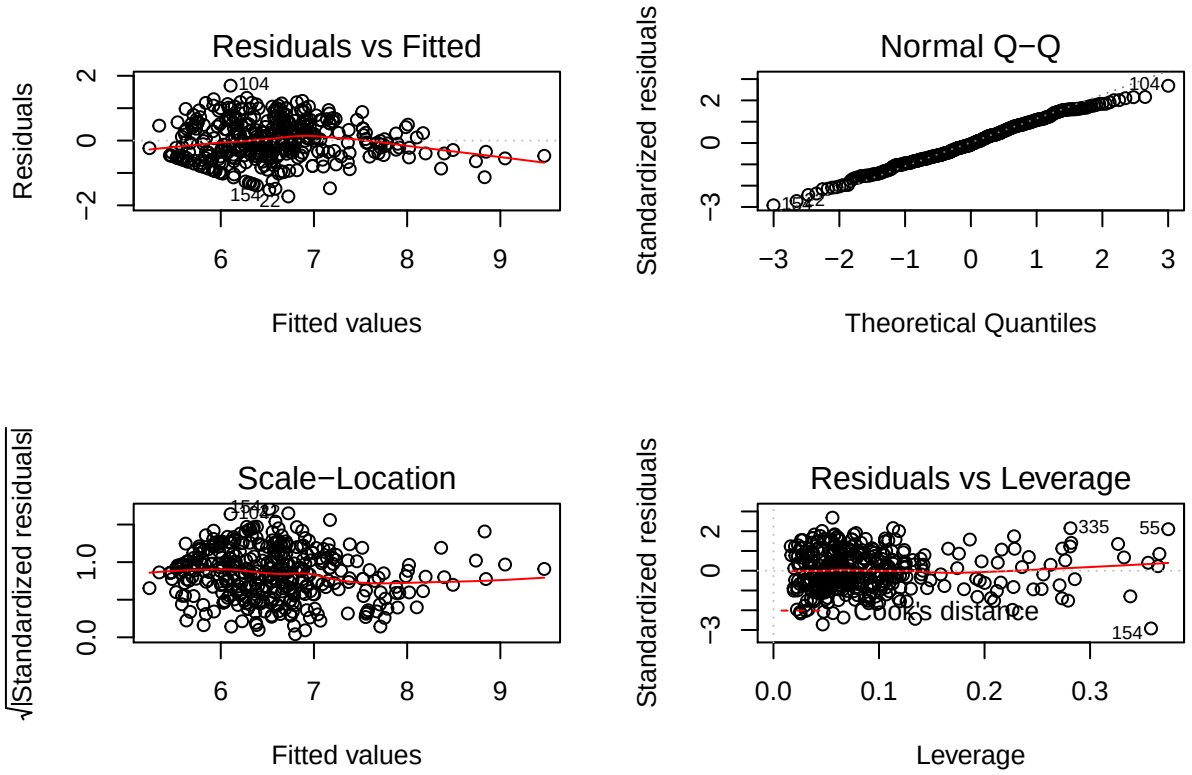
```
set.seed(123)
# check the imdb rating versus the testing prediction in Model 1
```

```
# they are lineary related which makes the testing confidential
pred <- predict(final.model, testing)
p <-ggplot(data = testing, aes(x=imdb_rating, y = pred)) + geom_point()
p + geom_abline(slope=1, intercept=0, colour = "red")
```



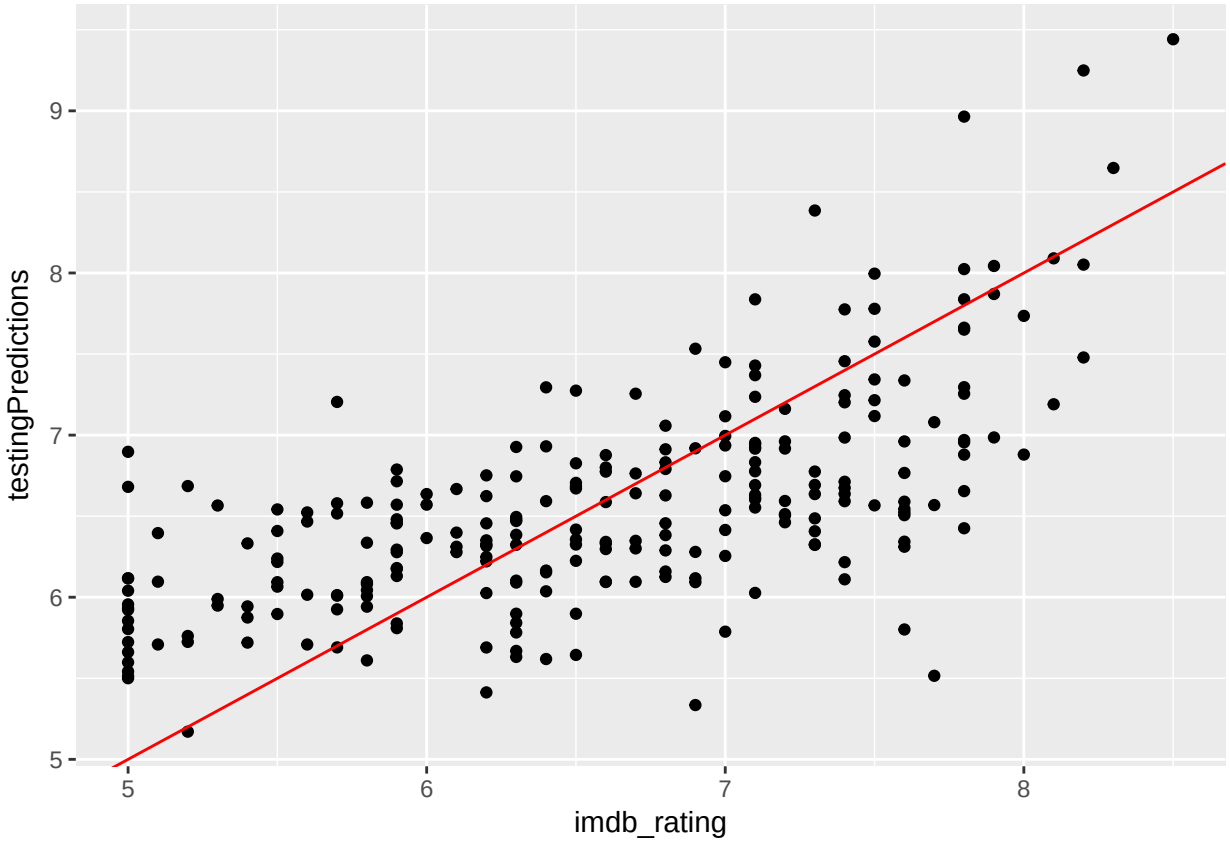
Plot 14 - Residuals and Diagnostics for Model 2

```
# check residuals and diagnostics for the final model in AIC
par(mfrow = c(2, 2))
plot(finalModEng)
```



Plot 15 - Predicted versus truth in test set for Model 2

```
set.seed(123)
# check the imdb rating versus the testing prediction in Model 2
# they are lineary related which makes the testing confidential
p <-ggplot(data = TestingPredictors, aes(x=imdb_rating, y = testingPredictions)) + geom_point()
p + geom_abline(slope=1, intercept=0, colour = "red")
```


testingPredictions
imdb_rating
5
6
7
8
9