Modelling and Prediction - What drives the success of a Movie?

Setup

Load packages

**library**(ggplot2)

**library**(dplyr)

**library**(GGally)

**library**(corrplot)

**library**(gridExtra)

Load data

**load**("movies.Rdata")

Part 1: Data

The raw data provided for this analysis consists of 651 movies which released prior to 2016. The data source comes from both IMDb and Rotten Tomatoes(RT) websites. Alongside with these 651 movies, there are 32 observation columns (i.e. variables) detailing the title of the movie, movie genre, runtime, MPAA rating, studio production, IMDB rating, RT audience rating, Oscar nominations, Actor/Actress/Director names and many more.

As this dataset consists of 651 **randomly selected** movies released prior to 2016. And based on the latest statistics in [IMDb website](http://www.imdb.com/stats), there are approximately 4 millions movie titles. In summary, the sample size is below 10% and may be relatively small to represent the population and also there was no information regarding about any specific sampling method used in this study.

Hence, this is more of an observational study instead of experimental. Any conclusions drawn can only be generalized to the population but cannot infer causality (as causality requires random assignment).

Part 2: Research question

As this movie dataset contained various characteristics information about the movie, it might be interesting to know how audiences rate the success of a movie. Do audience responded differently to different type of movies, the length of the movie, the year or month it’s released or more concerns over who casted in the movie? These information may provide some important insights about how popular the movie can be.

Popular movie is likely to attract bigger crowds which leading to higher ratings and hence generating more ticket sales which literally means better revenues to the production studio. In summary, every producers would wish to have better ratings for their movies as, to them this is the key driver to success.

My research question will assess if we can predict the success of a movie based on certain characteristics of a movie. A good prediction model would be valuable to the producers as they will know the likelihood of success of the movie before it is released to the public. In the dataset, there are three variables related to the scoring of a movie - IMDB rating vs two other variables from Rotten Tomatoes (i.e. Audience score/Critics score). In this analysis, I have pre-selected **IMDb rating** as the response/dependent variable\*.

\*Rationale as to why IMDb Rating as Response Variable

IMDb Rating vs Rotten Tomatoes Audience Scoring

IMDb is a mega movie database and has successfully become an industry standard resource and definitive reference for who-what-where-when information on movie production and personnel. The review and public-opinion features are really an afterthought to this primary purpose. Whilst, on the other hand, Rotten Tomatoes is a less-than-consistent approach to which critics and reviews are included. In conclusion, IMDb would have larger voters and generally accepted by many viewers across all different age groups.

How IMDb Rating and Rotten Tomatoes Audience Scoring are calculated?

IMDB weights the average based on how people fill out ratings and to avoid “vote stuffing”. In summary, it used a Weighted Average Ratings methodology. Rotten Tomatoes, on the other side, does not average the ratings. It counts the percentage of users that rate the movie at 3.5 stars or higher, the percentage that clearly think the movie is considered “good”.

Part 3: Exploratory data analysis

Analyse the structure of the dataset

Movie Dataset

**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/tt0205873/" "http://www.imdb.com/title/tt0118111/" "http://www.imdb.com/title/tt0106226/" ...

## $ rt\_url : chr "//www.rottentomatoes.com/m/filly\_brown\_2012/" "//www.rottentomatoes.com/m/dish/" "//www.rottentomatoes.com/m/waiting\_for\_guffman/" "//www.rottentomatoes.com/m/age\_of\_innocence/" ...

This function assesses the structure of each variables in the dataset. As you can seem there are  
1.Character Variables (9 columns)  
2.Factor Variables (12 columns)  
3.Numeric / Integer Variables (11 columns) - Of which 6 are date related

Before we proceed further, it is worthnoting if there are any missing values in the dataset. This is an important step when we are working with any regression model.

Remove Missing Values (NA)

Check<-**complete.cases**(movies)

dataset<-movies[Check,]

**dim**(dataset) *# Assess the number of records with full data values*

## [1] 619 32

We must ensure all missing values are excluded in the analysis as it may distort the regression model. In summary, we have excluded 32 records with missing values in our analysis. After ensuring a complete dataset with values, we will proceed to review and only include variables which are significant to the regression model.

Based on the dataset, column 25 to 32 are likely to be removed as these are related to actor/actress/director names and some URL links which may be too granular in this regression model. Also, this model predicts how successful the movie will be prior to its releasal. So column 10 to 12 will also be removed as these are DVD release dates which may not required to decide on the success of a movie.

Also, special note on **“studio”** variable as it has 211 levels and is too granular to be used in the regression model. Hence, this variable will be removed from the analysis. In a nutshell, we will deal with a smaller dataset in our next step.

Reduced Version - Movie Dataset

dataset<-dataset[**c**(1:5,7:9,13,16,18,14:15,17,19:24)] *#identify key variables which will be used in the model*

**str**(dataset) *# apply a second look on the dataset to ensure the data selected are correct*

## Classes 'tbl\_df', 'tbl' and 'data.frame': 619 obs. of 20 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 2 2 1 2 2 ...

## $ genre : Factor w/ 11 levels "Action & Adventure",..: 6 6 4 6 7 6 6 5 6 1 ...

## $ runtime : num 80 101 84 139 90 142 93 88 119 127 ...

## $ mpaa\_rating : Factor w/ 6 levels "G","NC-17","PG",..: 5 4 5 3 5 4 5 6 6 3 ...

## $ thtr\_rel\_year : num 2013 2001 1996 1993 2004 ...

## $ thtr\_rel\_month : num 4 3 8 10 9 1 11 9 3 6 ...

## $ thtr\_rel\_day : num 19 14 21 1 10 1 8 7 2 19 ...

## $ imdb\_rating : num 5.5 7.3 7.6 7.2 5.1 7.2 5.5 7.5 6.6 6.8 ...

## $ critics\_score : num 45 96 91 80 33 57 17 90 83 89 ...

## $ audience\_score : num 73 81 91 76 27 76 47 89 66 75 ...

## $ imdb\_num\_votes : int 899 12285 22381 35096 2386 5016 2272 880 12496 71979 ...

## $ critics\_rating : Factor w/ 3 levels "Certified Fresh",..: 3 1 1 1 3 3 3 2 1 1 ...

## $ audience\_rating : Factor w/ 2 levels "Spilled","Upright": 2 2 2 2 1 2 1 2 2 2 ...

## $ 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 2 1 1 2 ...

## $ 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 2 ...

**summary**(dataset)

## title title\_type genre

## Length:619 Documentary : 42 Drama :298

## Class :character Feature Film:573 Comedy : 86

## Mode :character TV Movie : 4 Action & Adventure: 62

## Mystery & Suspense: 56

## Documentary : 40

## Horror : 22

## (Other) : 55

## runtime mpaa\_rating thtr\_rel\_year thtr\_rel\_month

## Min. : 65.0 G : 16 Min. :1972 Min. : 1.000

## 1st Qu.: 93.0 NC-17 : 1 1st Qu.:1991 1st Qu.: 4.000

## Median :103.0 PG :111 Median :2000 Median : 7.000

## Mean :106.5 PG-13 :131 Mean :1998 Mean : 6.733

## 3rd Qu.:116.0 R :319 3rd Qu.:2007 3rd Qu.:10.000

## Max. :267.0 Unrated: 41 Max. :2014 Max. :12.000

##

## thtr\_rel\_day imdb\_rating critics\_score audience\_score

## Min. : 1.00 Min. :1.900 Min. : 1.00 Min. :11.00

## 1st Qu.: 7.00 1st Qu.:5.900 1st Qu.: 33.00 1st Qu.:46.00

## Median :15.00 Median :6.600 Median : 61.00 Median :65.00

## Mean :14.43 Mean :6.486 Mean : 57.43 Mean :62.21

## 3rd Qu.:22.00 3rd Qu.:7.300 3rd Qu.: 82.50 3rd Qu.:80.00

## Max. :31.00 Max. :9.000 Max. :100.00 Max. :97.00

##

## imdb\_num\_votes critics\_rating audience\_rating best\_pic\_nom

## Min. : 183 Certified Fresh:131 Spilled:264 no :597

## 1st Qu.: 5026 Fresh :195 Upright:355 yes: 22

## Median : 16480 Rotten :293

## Mean : 60014

## 3rd Qu.: 62507

## Max. :893008

##

## best\_pic\_win best\_actor\_win best\_actress\_win best\_dir\_win top200\_box

## no :612 no :528 no :548 no :576 no :604

## yes: 7 yes: 91 yes: 71 yes: 43 yes: 15

##

##

##

##

##

The other aspect of the work is to assess the data distribution of each movie characteristics (eg. by Title, by Genre, by RunTime, by MPAA Rating, by Theatre Release Year, Month or Day etc)

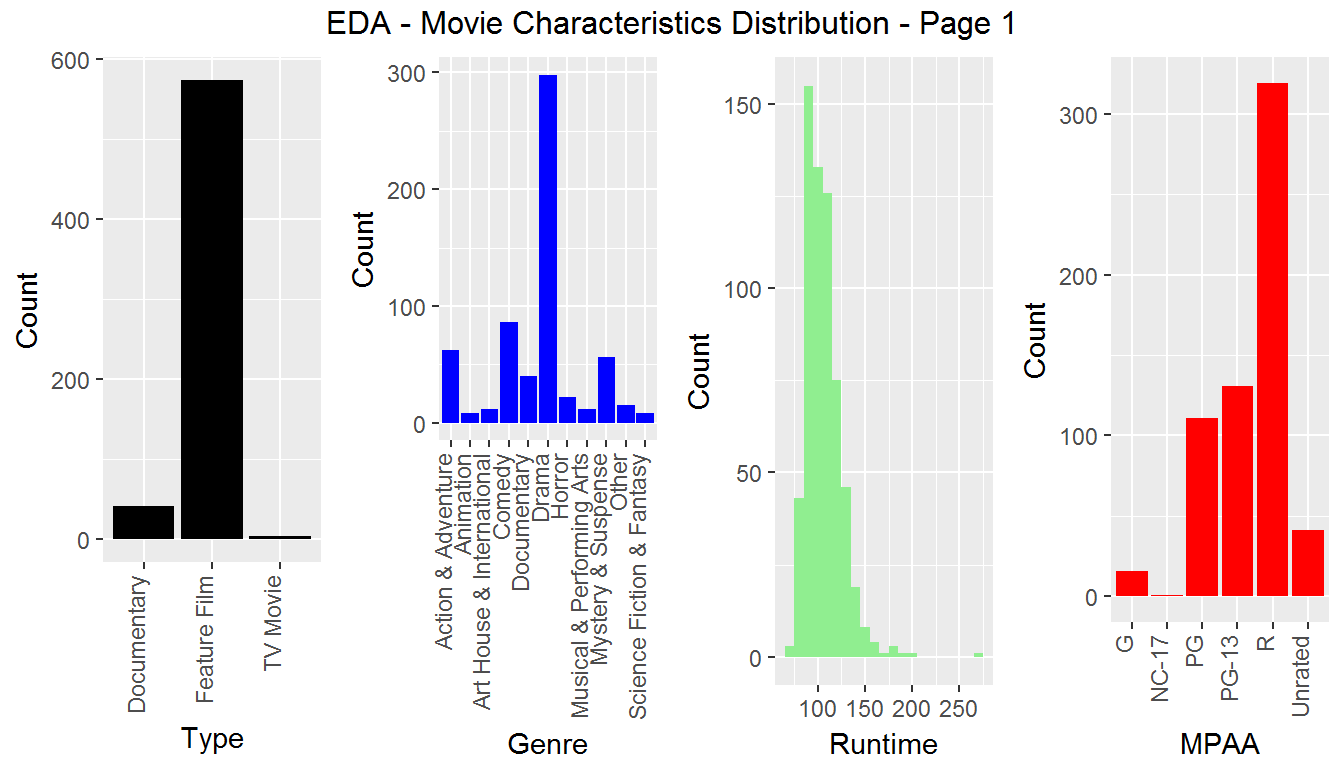
Type <-**ggplot**(data = dataset, **aes**(x = title\_type)) + **geom\_bar**(fill="black") + **xlab**("Type") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

Genre <-**ggplot**(data = dataset, **aes**(x = genre)) + **geom\_bar**(fill="blue") + **xlab**("Genre") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

RunTime <- **ggplot**(data = dataset, **aes**(x=runtime)) + **geom\_histogram**(binwidth=10,fill="light green") + **xlab**("Runtime") + **ylab**("Count")

MPAA <-**ggplot**(data = dataset, **aes**(x = mpaa\_rating)) + **geom\_bar**(fill="red") + **xlab**("MPAA") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

**grid.arrange**(Type,Genre,RunTime,MPAA,nrow=1,top="EDA - Movie Characteristics Distribution - Page 1")

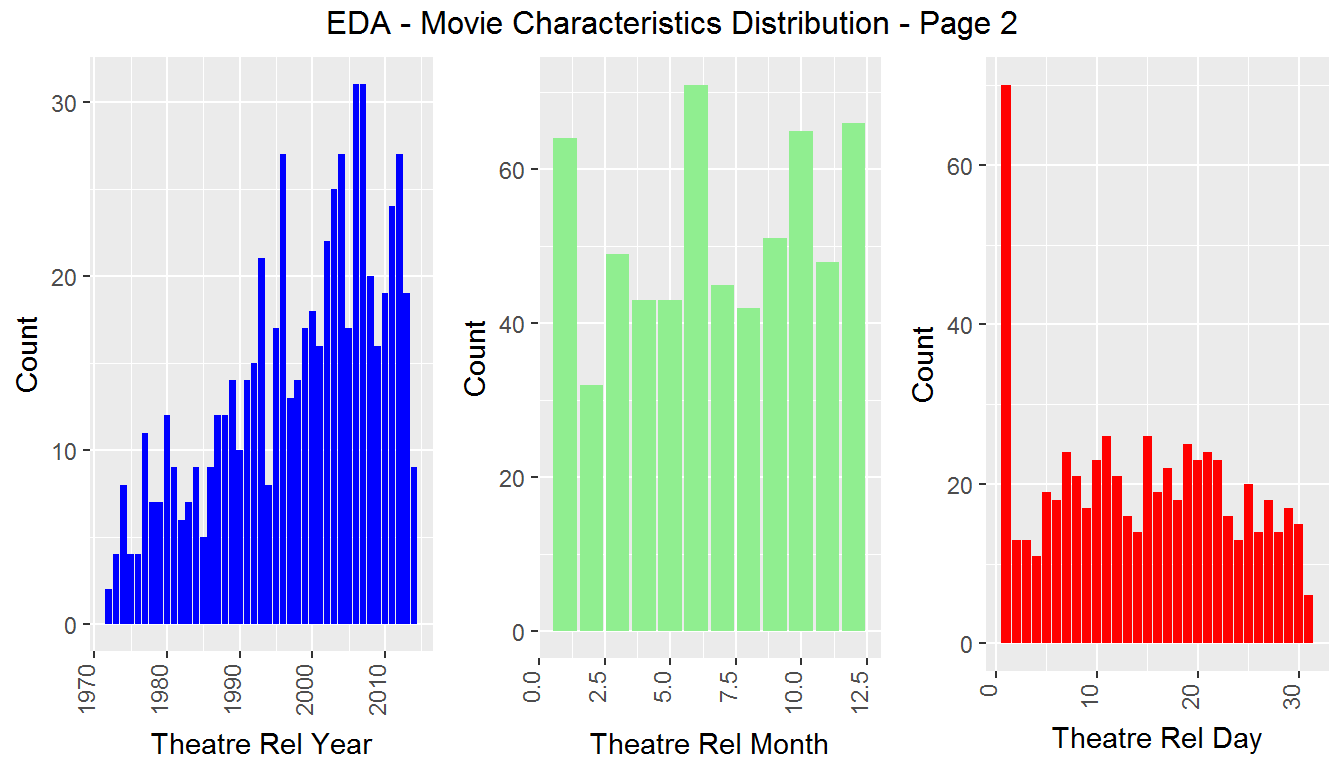


Thtr\_Yr <-**ggplot**(data = dataset, **aes**(x = thtr\_rel\_year)) + **geom\_bar**(fill="blue") + **xlab**("Theatre Rel Year") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

Thtr\_Month <-**ggplot**(data = dataset, **aes**(x = thtr\_rel\_month)) + **geom\_bar**(fill="light green") + **xlab**("Theatre Rel Month") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

Thtr\_Day <-**ggplot**(data = dataset, **aes**(x = thtr\_rel\_day)) + **geom\_bar**(fill="red") + **xlab**("Theatre Rel Day") + **ylab**("Count") + **theme**(axis.text.x=**element\_text**(angle=90, hjust=1, vjust=0))

**grid.arrange**(Thtr\_Yr,Thtr\_Month,Thtr\_Day, nrow=1,top="EDA - Movie Characteristics Distribution - Page 2")



After reviewed the distribution charts above, we will remain with 619 movies data for regression analysis.

Part 4: Modeling

Check Collinearity

After explored on the data, the next thing is to check if there is any collinearity within the **numerical explanatory** variables. To do this, I have created a sub-dataset of all the numeric variables e.g. runtime, thtr\_rel\_year, thtr\_rel\_month, thtr\_rel\_day, imdb\_num\_votes, critics\_score, audience\_score. This should ease us to do a correlation matrix to understand each of these explanatory variables.

num\_expl\_var <- dataset[**c**(4,6:8,10:12)] *#identify those numerical explanatory variables*

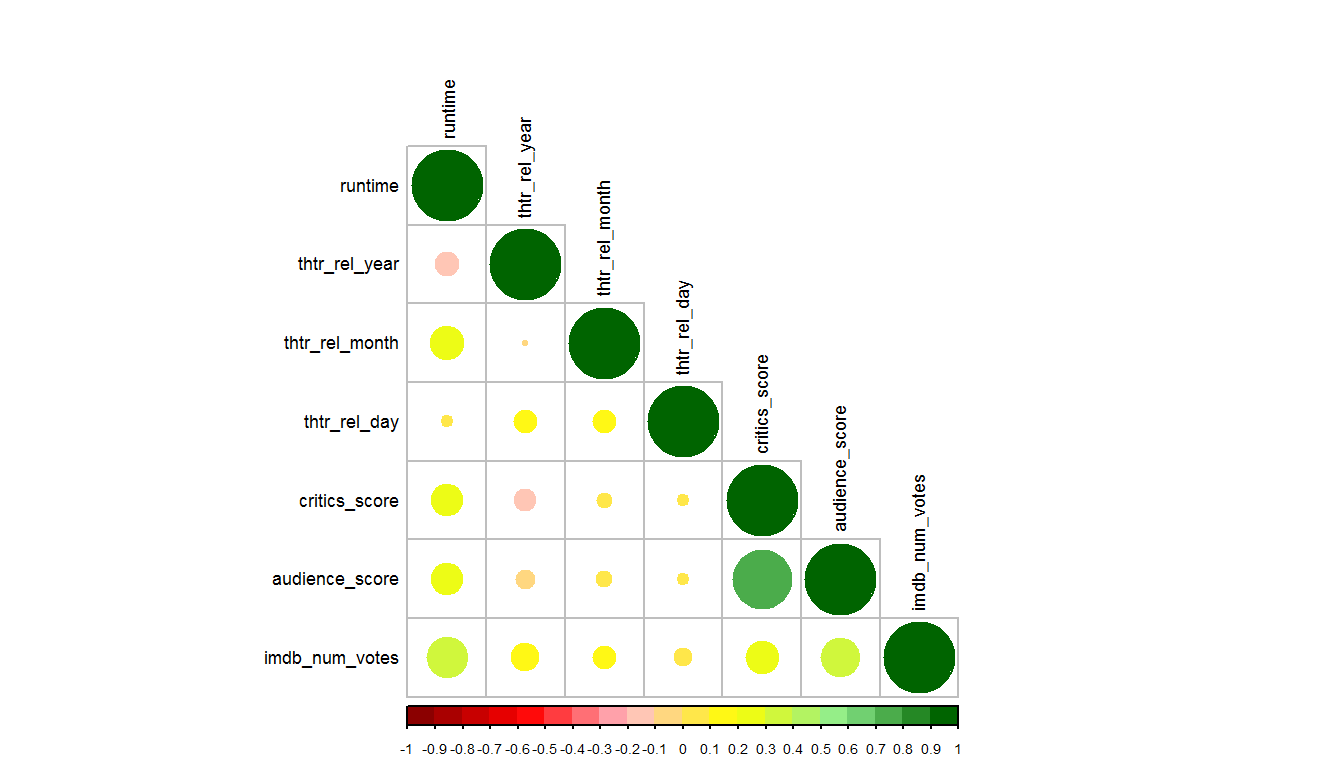
corr<- **cor**(num\_expl\_var)

cex.before <- **par**("cex")

**par**(cex = 0.55)

col<- **colorRampPalette**(**c**("dark red","red","pink", "yellow","light green", "dark green"))(20)

**corrplot**(corr, method="circle", type="lower", col=col, sig.level = 0.01, tl.col="black")

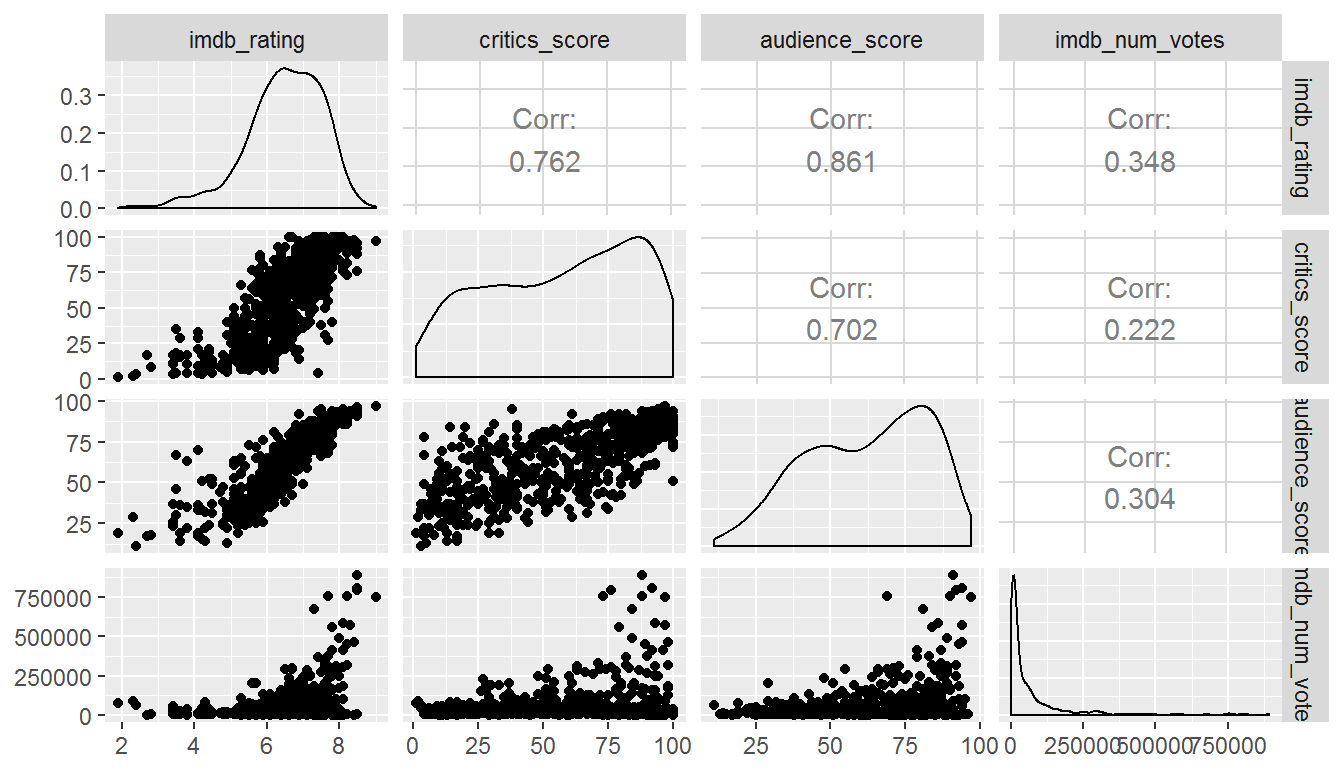


In the correlogram chart above, it seems Critics\_score and Audience\_score are highly correlated (i.e.collinear). Literally it means if the two variables are added as explanatory variables, this will distort and complicate the model and hence redundant. As a result, one of these variables has to be removed from the model. In this case, I have chosen to remove critics\_score from my regression model.

Why Critics Score and not Audience Score?

To address that, we have worked out the correlation coefficents between these explanatory variables with response variable. Audience score seems to be more correlated to the response variable and also audience scoring is derived from a larger group than critics score. Hence, audience score is selected over critics score.

**ggpairs**(dataset,columns=9:12)



Create Multiple Regression Model - using Backward Stepwise Regression approach

Firstly, we will create a full model and then work backward by eliminating variables which has the highest p-value. This process will take a while before we reach our final model. This approach is called backward stepwise regression.

model <- dataset[**c**(2:9,11:20)]

first\_mod <-**lm**(imdb\_rating~., data=model)

**summary**(first\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ ., data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.5391 -0.1681 0.0466 0.2527 1.0848

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 4.047e+00 4.341e+00 0.932 0.351550

## title\_typeFeature Film -3.515e-01 1.939e-01 -1.812 0.070439 .

## title\_typeTV Movie -3.271e-01 3.110e-01 -1.052 0.293339

## genreAnimation -5.611e-01 2.005e-01 -2.798 0.005315 \*\*

## genreArt House & International 3.368e-01 1.593e-01 2.114 0.034915 \*

## genreComedy -1.437e-01 8.269e-02 -1.738 0.082744 .

## genreDocumentary 1.157e-01 2.053e-01 0.563 0.573329

## genreDrama 1.420e-01 7.300e-02 1.945 0.052260 .

## genreHorror 1.153e-01 1.241e-01 0.930 0.352968

## genreMusical & Performing Arts 5.785e-02 1.691e-01 0.342 0.732475

## genreMystery & Suspense 2.853e-01 9.354e-02 3.050 0.002388 \*\*

## genreOther 5.667e-02 1.429e-01 0.396 0.691914

## genreScience Fiction & Fantasy -7.923e-02 1.827e-01 -0.434 0.664733

## runtime 4.307e-03 1.271e-03 3.389 0.000747 \*\*\*

## mpaa\_ratingNC-17 9.980e-02 5.058e-01 0.197 0.843640

## mpaa\_ratingPG -1.580e-01 1.434e-01 -1.102 0.271029

## mpaa\_ratingPG-13 -1.761e-01 1.493e-01 -1.180 0.238644

## mpaa\_ratingR -1.090e-01 1.441e-01 -0.757 0.449579

## mpaa\_ratingUnrated -1.672e-01 1.717e-01 -0.974 0.330701

## thtr\_rel\_year -1.144e-04 2.158e-03 -0.053 0.957739

## thtr\_rel\_month 9.412e-03 5.849e-03 1.609 0.108115

## thtr\_rel\_day -1.000e-03 2.257e-03 -0.443 0.657910

## audience\_score 4.616e-02 2.180e-03 21.173 < 2e-16 \*\*\*

## imdb\_num\_votes 7.955e-07 2.321e-07 3.427 0.000652 \*\*\*

## critics\_ratingFresh -3.507e-02 6.299e-02 -0.557 0.577961

## critics\_ratingRotten -3.019e-01 6.741e-02 -4.478 9.05e-06 \*\*\*

## audience\_ratingUpright -4.243e-01 7.947e-02 -5.339 1.34e-07 \*\*\*

## best\_pic\_nomyes -1.075e-01 1.289e-01 -0.834 0.404418

## best\_pic\_winyes -3.971e-02 2.249e-01 -0.177 0.859895

## best\_actor\_winyes 2.219e-02 5.842e-02 0.380 0.704198

## best\_actress\_winyes 8.711e-02 6.435e-02 1.354 0.176339

## best\_dir\_winyes 5.314e-02 8.392e-02 0.633 0.526866

## top200\_boxyes -9.827e-02 1.367e-01 -0.719 0.472376

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4816 on 586 degrees of freedom

## Multiple R-squared: 0.8096, Adjusted R-squared: 0.7992

## F-statistic: 77.88 on 32 and 586 DF, p-value: < 2.2e-16

**anova**(first\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 145.3158 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.1252 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 154.8520 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6211 1.361e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2388 0.03995 \*

## thtr\_rel\_month 1 0.66 0.66 2.8573 0.09149 .

## thtr\_rel\_day 1 0.42 0.42 1.8061 0.17949

## audience\_score 1 344.24 344.24 1484.4354 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.1951 8.429e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.4689 2.840e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.3277 8.930e-08 \*\*\*

## best\_pic\_nom 1 0.12 0.12 0.5253 0.46889

## best\_pic\_win 1 0.00 0.00 0.0015 0.96878

## best\_actor\_win 1 0.05 0.05 0.2177 0.64096

## best\_actress\_win 1 0.41 0.41 1.7553 0.18573

## best\_dir\_win 1 0.10 0.10 0.4305 0.51202

## top200\_box 1 0.12 0.12 0.5171 0.47238

## Residuals 586 135.89 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

This probably a good start as the adj R2 of 0.7992 is relatively strong. But the next procedure is to remove those insignificant variables one-by-one, based on the highest p-value first. In this specific model, **best\_pic\_win** has the highest p-value and hence will be removed accordingly. This is an iterative process and may take some time to get to the final model.

second\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+thtr\_rel\_day+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating+best\_pic\_nom+best\_actor\_win+

best\_actress\_win+best\_dir\_win, data=model)

**summary**(second\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + thtr\_rel\_day + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating + best\_pic\_nom +

## best\_actor\_win + best\_actress\_win + best\_dir\_win, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.5427 -0.1683 0.0389 0.2524 1.0862

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.719e+00 4.309e+00 0.863 0.388495

## title\_typeFeature Film -3.523e-01 1.937e-01 -1.819 0.069420 .

## title\_typeTV Movie -3.302e-01 3.106e-01 -1.063 0.288151

## genreAnimation -5.474e-01 1.993e-01 -2.746 0.006213 \*\*

## genreArt House & International 3.401e-01 1.590e-01 2.140 0.032794 \*

## genreComedy -1.400e-01 8.225e-02 -1.703 0.089184 .

## genreDocumentary 1.182e-01 2.050e-01 0.576 0.564518

## genreDrama 1.458e-01 7.264e-02 2.006 0.045261 \*

## genreHorror 1.184e-01 1.238e-01 0.956 0.339465

## genreMusical & Performing Arts 6.193e-02 1.688e-01 0.367 0.713808

## genreMystery & Suspense 2.891e-01 9.322e-02 3.101 0.002019 \*\*

## genreOther 6.195e-02 1.424e-01 0.435 0.663635

## genreScience Fiction & Fantasy -8.187e-02 1.824e-01 -0.449 0.653750

## runtime 4.283e-03 1.269e-03 3.376 0.000785 \*\*\*

## mpaa\_ratingNC-17 1.169e-01 5.046e-01 0.232 0.816914

## mpaa\_ratingPG -1.494e-01 1.427e-01 -1.047 0.295512

## mpaa\_ratingPG-13 -1.644e-01 1.483e-01 -1.109 0.267989

## mpaa\_ratingR -9.556e-02 1.427e-01 -0.669 0.503442

## mpaa\_ratingUnrated -1.568e-01 1.708e-01 -0.918 0.359183

## thtr\_rel\_year 4.103e-05 2.143e-03 0.019 0.984733

## thtr\_rel\_month 9.214e-03 5.819e-03 1.583 0.113853

## thtr\_rel\_day -9.823e-04 2.253e-03 -0.436 0.662996

## audience\_score 4.626e-02 2.171e-03 21.307 < 2e-16 \*\*\*

## imdb\_num\_votes 7.524e-07 2.228e-07 3.377 0.000780 \*\*\*

## critics\_ratingFresh -3.087e-02 6.264e-02 -0.493 0.622322

## critics\_ratingRotten -2.977e-01 6.709e-02 -4.437 1.09e-05 \*\*\*

## audience\_ratingUpright -4.271e-01 7.928e-02 -5.387 1.03e-07 \*\*\*

## best\_pic\_nomyes -1.126e-01 1.179e-01 -0.955 0.339960

## best\_actor\_winyes 2.183e-02 5.819e-02 0.375 0.707655

## best\_actress\_winyes 8.454e-02 6.415e-02 1.318 0.188054

## best\_dir\_winyes 5.105e-02 8.052e-02 0.634 0.526354

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.481 on 588 degrees of freedom

## Multiple R-squared: 0.8094, Adjusted R-squared: 0.7997

## F-statistic: 83.26 on 30 and 588 DF, p-value: < 2.2e-16

**anova**(second\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 145.6759 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.2246 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.2357 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6548 1.259e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2493 0.03971 \*

## thtr\_rel\_month 1 0.66 0.66 2.8644 0.09109 .

## thtr\_rel\_day 1 0.42 0.42 1.8106 0.17896

## audience\_score 1 344.24 344.24 1488.1133 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.2451 8.213e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.5073 2.735e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.4003 8.606e-08 \*\*\*

## best\_pic\_nom 1 0.12 0.12 0.5266 0.46834

## best\_actor\_win 1 0.05 0.05 0.2153 0.64281

## best\_actress\_win 1 0.41 0.41 1.7641 0.18463

## best\_dir\_win 1 0.09 0.09 0.4019 0.52635

## Residuals 588 136.02 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The adj R2 has improved but the model still has some insignificant predictors. We will continue to eliminate the insignificant predictors one-by-one. The highest p value in the second model is **best\_actor\_win** variable.

third\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+thtr\_rel\_day+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating+best\_pic\_nom+

best\_actress\_win+best\_dir\_win, data=model)

**summary**(third\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + thtr\_rel\_day + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating + best\_pic\_nom +

## best\_actress\_win + best\_dir\_win, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54348 -0.16429 0.03922 0.25058 1.08392

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.709e+00 4.306e+00 0.861 0.389437

## title\_typeFeature Film -3.510e-01 1.935e-01 -1.814 0.070172 .

## title\_typeTV Movie -3.318e-01 3.104e-01 -1.069 0.285437

## genreAnimation -5.456e-01 1.991e-01 -2.740 0.006327 \*\*

## genreArt House & International 3.379e-01 1.587e-01 2.129 0.033706 \*

## genreComedy -1.402e-01 8.219e-02 -1.705 0.088653 .

## genreDocumentary 1.178e-01 2.048e-01 0.575 0.565341

## genreDrama 1.464e-01 7.257e-02 2.018 0.044068 \*

## genreHorror 1.169e-01 1.237e-01 0.945 0.344967

## genreMusical & Performing Arts 6.077e-02 1.686e-01 0.360 0.718691

## genreMystery & Suspense 2.921e-01 9.280e-02 3.148 0.001728 \*\*

## genreOther 6.263e-02 1.423e-01 0.440 0.659933

## genreScience Fiction & Fantasy -8.394e-02 1.822e-01 -0.461 0.645200

## runtime 4.367e-03 1.248e-03 3.499 0.000502 \*\*\*

## mpaa\_ratingNC-17 1.164e-01 5.042e-01 0.231 0.817450

## mpaa\_ratingPG -1.478e-01 1.426e-01 -1.037 0.300153

## mpaa\_ratingPG-13 -1.637e-01 1.482e-01 -1.105 0.269627

## mpaa\_ratingR -9.508e-02 1.426e-01 -0.667 0.505272

## mpaa\_ratingUnrated -1.566e-01 1.707e-01 -0.917 0.359406

## thtr\_rel\_year 4.112e-05 2.141e-03 0.019 0.984687

## thtr\_rel\_month 9.321e-03 5.808e-03 1.605 0.109038

## thtr\_rel\_day -9.664e-04 2.251e-03 -0.429 0.667841

## audience\_score 4.628e-02 2.169e-03 21.342 < 2e-16 \*\*\*

## imdb\_num\_votes 7.491e-07 2.224e-07 3.368 0.000807 \*\*\*

## critics\_ratingFresh -3.038e-02 6.258e-02 -0.486 0.627500

## critics\_ratingRotten -2.976e-01 6.704e-02 -4.439 1.08e-05 \*\*\*

## audience\_ratingUpright -4.285e-01 7.912e-02 -5.416 8.89e-08 \*\*\*

## best\_pic\_nomyes -1.087e-01 1.173e-01 -0.927 0.354523

## best\_actress\_winyes 8.590e-02 6.400e-02 1.342 0.180049

## best\_dir\_winyes 5.165e-02 8.045e-02 0.642 0.521138

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4806 on 589 degrees of freedom

## Multiple R-squared: 0.8094, Adjusted R-squared: 0.8

## F-statistic: 86.25 on 29 and 589 DF, p-value: < 2.2e-16

**anova**(third\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 145.8887 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.2833 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.4625 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6748 1.203e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2555 0.03956 \*

## thtr\_rel\_month 1 0.66 0.66 2.8686 0.09085 .

## thtr\_rel\_day 1 0.42 0.42 1.8132 0.17864

## audience\_score 1 344.24 344.24 1490.2874 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.2747 8.089e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.5299 2.675e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.4433 8.421e-08 \*\*\*

## best\_pic\_nom 1 0.12 0.12 0.5273 0.46802

## best\_actress\_win 1 0.42 0.42 1.8316 0.17645

## best\_dir\_win 1 0.10 0.10 0.4121 0.52114

## Residuals 589 136.05 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The adj R2 has improved to 0.8. The next variable to be eliminated is **best\_dir\_win**.

fourth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+thtr\_rel\_day+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating+best\_pic\_nom+

best\_actress\_win, data=model)

**summary**(fourth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + thtr\_rel\_day + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating + best\_pic\_nom +

## best\_actress\_win, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54251 -0.16980 0.03898 0.25007 1.07756

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.898e+00 4.294e+00 0.908 0.364422

## title\_typeFeature Film -3.473e-01 1.933e-01 -1.797 0.072894 .

## title\_typeTV Movie -3.299e-01 3.102e-01 -1.064 0.287919

## genreAnimation -5.446e-01 1.990e-01 -2.737 0.006396 \*\*

## genreArt House & International 3.352e-01 1.586e-01 2.114 0.034959 \*

## genreComedy -1.400e-01 8.215e-02 -1.704 0.088952 .

## genreDocumentary 1.187e-01 2.047e-01 0.580 0.562240

## genreDrama 1.455e-01 7.252e-02 2.007 0.045204 \*

## genreHorror 1.173e-01 1.236e-01 0.949 0.342819

## genreMusical & Performing Arts 6.113e-02 1.685e-01 0.363 0.716988

## genreMystery & Suspense 2.924e-01 9.276e-02 3.152 0.001703 \*\*

## genreOther 5.847e-02 1.421e-01 0.412 0.680797

## genreScience Fiction & Fantasy -8.223e-02 1.821e-01 -0.452 0.651762

## runtime 4.493e-03 1.232e-03 3.647 0.000289 \*\*\*

## mpaa\_ratingNC-17 1.183e-01 5.040e-01 0.235 0.814531

## mpaa\_ratingPG -1.439e-01 1.424e-01 -1.011 0.312445

## mpaa\_ratingPG-13 -1.608e-01 1.480e-01 -1.086 0.277748

## mpaa\_ratingR -9.120e-02 1.424e-01 -0.640 0.522205

## mpaa\_ratingUnrated -1.547e-01 1.706e-01 -0.907 0.364890

## thtr\_rel\_year -6.205e-05 2.134e-03 -0.029 0.976818

## thtr\_rel\_month 9.375e-03 5.804e-03 1.615 0.106806

## thtr\_rel\_day -9.780e-04 2.250e-03 -0.435 0.663947

## audience\_score 4.633e-02 2.166e-03 21.386 < 2e-16 \*\*\*

## imdb\_num\_votes 7.599e-07 2.217e-07 3.428 0.000651 \*\*\*

## critics\_ratingFresh -3.019e-02 6.255e-02 -0.483 0.629565

## critics\_ratingRotten -3.003e-01 6.686e-02 -4.492 8.49e-06 \*\*\*

## audience\_ratingUpright -4.312e-01 7.898e-02 -5.459 7.05e-08 \*\*\*

## best\_pic\_nomyes -1.052e-01 1.171e-01 -0.898 0.369509

## best\_actress\_winyes 8.660e-02 6.396e-02 1.354 0.176236

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4804 on 590 degrees of freedom

## Multiple R-squared: 0.8093, Adjusted R-squared: 0.8002

## F-statistic: 89.4 on 28 and 590 DF, p-value: < 2.2e-16

**anova**(fourth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.0342 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.3235 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.6175 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6884 1.166e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2598 0.03946 \*

## thtr\_rel\_month 1 0.66 0.66 2.8715 0.09069 .

## thtr\_rel\_day 1 0.42 0.42 1.8150 0.17842

## audience\_score 1 344.24 344.24 1491.7737 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.2949 8.004e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.5454 2.634e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.4727 8.295e-08 \*\*\*

## best\_pic\_nom 1 0.12 0.12 0.5279 0.46779

## best\_actress\_win 1 0.42 0.42 1.8335 0.17624

## Residuals 590 136.15 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The highest p value above is **best\_pic\_nom** variable. We will remove this variable in our regression.

fifth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+thtr\_rel\_day+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating+

best\_actress\_win, data=model)

**summary**(fifth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + thtr\_rel\_day + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating + best\_actress\_win,

## data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54575 -0.17002 0.03634 0.25551 1.07724

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.707e+00 4.288e+00 0.865 0.387641

## title\_typeFeature Film -3.490e-01 1.933e-01 -1.806 0.071503 .

## title\_typeTV Movie -3.335e-01 3.101e-01 -1.075 0.282592

## genreAnimation -5.454e-01 1.990e-01 -2.741 0.006309 \*\*

## genreArt House & International 3.338e-01 1.586e-01 2.105 0.035719 \*

## genreComedy -1.423e-01 8.209e-02 -1.734 0.083452 .

## genreDocumentary 1.169e-01 2.047e-01 0.571 0.568027

## genreDrama 1.420e-01 7.240e-02 1.962 0.050256 .

## genreHorror 1.126e-01 1.235e-01 0.912 0.362291

## genreMusical & Performing Arts 6.323e-02 1.685e-01 0.375 0.707624

## genreMystery & Suspense 2.894e-01 9.268e-02 3.122 0.001881 \*\*

## genreOther 5.015e-02 1.417e-01 0.354 0.723574

## genreScience Fiction & Fantasy -8.128e-02 1.821e-01 -0.446 0.655447

## runtime 4.388e-03 1.226e-03 3.578 0.000374 \*\*\*

## mpaa\_ratingNC-17 1.241e-01 5.038e-01 0.246 0.805524

## mpaa\_ratingPG -1.471e-01 1.423e-01 -1.034 0.301746

## mpaa\_ratingPG-13 -1.637e-01 1.480e-01 -1.106 0.268974

## mpaa\_ratingR -9.207e-02 1.424e-01 -0.647 0.518178

## mpaa\_ratingUnrated -1.546e-01 1.706e-01 -0.906 0.365172

## thtr\_rel\_year 4.435e-05 2.131e-03 0.021 0.983399

## thtr\_rel\_month 8.720e-03 5.757e-03 1.515 0.130402

## thtr\_rel\_day -9.169e-04 2.248e-03 -0.408 0.683571

## audience\_score 4.617e-02 2.159e-03 21.388 < 2e-16 \*\*\*

## imdb\_num\_votes 7.296e-07 2.191e-07 3.330 0.000921 \*\*\*

## critics\_ratingFresh -2.333e-02 6.207e-02 -0.376 0.707183

## critics\_ratingRotten -2.938e-01 6.646e-02 -4.421 1.17e-05 \*\*\*

## audience\_ratingUpright -4.276e-01 7.886e-02 -5.422 8.61e-08 \*\*\*

## best\_actress\_winyes 7.906e-02 6.339e-02 1.247 0.212848

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4803 on 591 degrees of freedom

## Multiple R-squared: 0.809, Adjusted R-squared: 0.8003

## F-statistic: 92.72 on 27 and 591 DF, p-value: < 2.2e-16

**anova**(fifth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.0820 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.3367 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.6685 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6929 1.151e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2612 0.03943 \*

## thtr\_rel\_month 1 0.66 0.66 2.8724 0.09064 .

## thtr\_rel\_day 1 0.42 0.42 1.8156 0.17835

## audience\_score 1 344.24 344.24 1492.2622 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.3016 7.975e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.5505 2.619e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.4823 8.251e-08 \*\*\*

## best\_actress\_win 1 0.36 0.36 1.5553 0.21285

## Residuals 591 136.33 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The predictor with highest p value is **best\_actress\_win**. By now, you would notice those variables relating to Oscar like Oscar nomination, Oscar winning or best actor/actress/director winners have little or no significance in contributing to the final model. And by removing them, we can see the adj R2 is getting stronger.

sixth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+thtr\_rel\_day+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating, data=model)

**summary**(sixth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + thtr\_rel\_day + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54920 -0.17859 0.03308 0.25704 1.07702

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.933e+00 4.286e+00 0.918 0.359174

## title\_typeFeature Film -3.490e-01 1.934e-01 -1.805 0.071604 .

## title\_typeTV Movie -3.203e-01 3.101e-01 -1.033 0.302092

## genreAnimation -5.323e-01 1.988e-01 -2.678 0.007620 \*\*

## genreArt House & International 3.435e-01 1.584e-01 2.168 0.030557 \*

## genreComedy -1.320e-01 8.171e-02 -1.616 0.106720

## genreDocumentary 1.241e-01 2.047e-01 0.606 0.544576

## genreDrama 1.546e-01 7.173e-02 2.156 0.031515 \*

## genreHorror 1.165e-01 1.235e-01 0.944 0.345663

## genreMusical & Performing Arts 6.406e-02 1.686e-01 0.380 0.704105

## genreMystery & Suspense 3.042e-01 9.196e-02 3.309 0.000995 \*\*\*

## genreOther 5.892e-02 1.416e-01 0.416 0.677545

## genreScience Fiction & Fantasy -8.064e-02 1.822e-01 -0.443 0.658151

## runtime 4.589e-03 1.216e-03 3.774 0.000177 \*\*\*

## mpaa\_ratingNC-17 1.115e-01 5.040e-01 0.221 0.825043

## mpaa\_ratingPG -1.450e-01 1.423e-01 -1.018 0.308901

## mpaa\_ratingPG-13 -1.628e-01 1.480e-01 -1.100 0.271811

## mpaa\_ratingR -9.354e-02 1.425e-01 -0.657 0.511692

## mpaa\_ratingUnrated -1.565e-01 1.707e-01 -0.917 0.359489

## thtr\_rel\_year -7.682e-05 2.130e-03 -0.036 0.971235

## thtr\_rel\_month 8.783e-03 5.760e-03 1.525 0.127819

## thtr\_rel\_day -8.508e-04 2.249e-03 -0.378 0.705311

## audience\_score 4.611e-02 2.159e-03 21.358 < 2e-16 \*\*\*

## imdb\_num\_votes 7.401e-07 2.190e-07 3.379 0.000775 \*\*\*

## critics\_ratingFresh -3.021e-02 6.186e-02 -0.488 0.625476

## critics\_ratingRotten -2.986e-01 6.638e-02 -4.498 8.25e-06 \*\*\*

## audience\_ratingUpright -4.282e-01 7.890e-02 -5.427 8.36e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4805 on 592 degrees of freedom

## Multiple R-squared: 0.8085, Adjusted R-squared: 0.8001

## F-statistic: 96.13 on 26 and 592 DF, p-value: < 2.2e-16

**anova**(sixth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 145.9451 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.2989 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.5226 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6801 1.180e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2572 0.03952 \*

## thtr\_rel\_month 1 0.66 0.66 2.8697 0.09079 .

## thtr\_rel\_day 1 0.42 0.42 1.8139 0.17855

## audience\_score 1 344.24 344.24 1490.8638 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.68 4.68 20.2825 8.050e-06 \*\*\*

## critics\_rating 2 7.17 3.59 15.5359 2.654e-07 \*\*\*

## audience\_rating 1 6.80 6.80 29.4547 8.358e-08 \*\*\*

## Residuals 592 136.69 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

It seems the adj R2 has decreased slightly but insignificant. But as my approach is using p-value elimination process and there are still one of two insignificant predictors (i.e. **thtr\_rel\_day & thtr\_rel\_month**), I will continue to remove them and assess the overall impact again. In my next iterative process, I will be removing **thtr\_rel\_day** variable as it has the highest p-value.

seventh\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+thtr\_rel\_month+

audience\_score+imdb\_num\_votes+critics\_rating+audience\_rating, data=model)

**summary**(seventh\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + thtr\_rel\_month + audience\_score + imdb\_num\_votes +

## critics\_rating + audience\_rating, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.55747 -0.17950 0.03493 0.25402 1.08478

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 4.038e+00 4.274e+00 0.945 0.345183

## title\_typeFeature Film -3.487e-01 1.932e-01 -1.804 0.071680 .

## title\_typeTV Movie -3.236e-01 3.097e-01 -1.045 0.296526

## genreAnimation -5.314e-01 1.986e-01 -2.675 0.007676 \*\*

## genreArt House & International 3.394e-01 1.580e-01 2.149 0.032051 \*

## genreComedy -1.311e-01 8.162e-02 -1.606 0.108700

## genreDocumentary 1.235e-01 2.045e-01 0.604 0.546126

## genreDrama 1.540e-01 7.166e-02 2.149 0.032053 \*

## genreHorror 1.140e-01 1.232e-01 0.925 0.355392

## genreMusical & Performing Arts 6.305e-02 1.684e-01 0.374 0.708283

## genreMystery & Suspense 3.028e-01 9.181e-02 3.298 0.001033 \*\*

## genreOther 6.076e-02 1.414e-01 0.430 0.667623

## genreScience Fiction & Fantasy -7.643e-02 1.817e-01 -0.421 0.674121

## runtime 4.596e-03 1.215e-03 3.783 0.000171 \*\*\*

## mpaa\_ratingNC-17 1.051e-01 5.033e-01 0.209 0.834740

## mpaa\_ratingPG -1.457e-01 1.422e-01 -1.025 0.305917

## mpaa\_ratingPG-13 -1.646e-01 1.478e-01 -1.113 0.266125

## mpaa\_ratingR -9.307e-02 1.424e-01 -0.654 0.513517

## mpaa\_ratingUnrated -1.546e-01 1.705e-01 -0.907 0.364889

## thtr\_rel\_year -1.357e-04 2.122e-03 -0.064 0.949021

## thtr\_rel\_month 8.557e-03 5.725e-03 1.495 0.135506

## audience\_score 4.612e-02 2.157e-03 21.379 < 2e-16 \*\*\*

## imdb\_num\_votes 7.400e-07 2.189e-07 3.381 0.000769 \*\*\*

## critics\_ratingFresh -2.788e-02 6.151e-02 -0.453 0.650467

## critics\_ratingRotten -2.968e-01 6.616e-02 -4.486 8.73e-06 \*\*\*

## audience\_ratingUpright -4.288e-01 7.883e-02 -5.440 7.80e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4802 on 593 degrees of freedom

## Multiple R-squared: 0.8085, Adjusted R-squared: 0.8004

## F-statistic: 100.1 on 25 and 593 DF, p-value: < 2.2e-16

**anova**(seventh\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.1563 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.3572 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.7477 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6998 1.128e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2633 0.03938 \*

## thtr\_rel\_month 1 0.66 0.66 2.8739 0.09055 .

## audience\_score 1 344.64 344.64 1494.7571 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.66 4.66 20.2106 8.345e-06 \*\*\*

## critics\_rating 2 7.16 3.58 15.5289 2.670e-07 \*\*\*

## audience\_rating 1 6.82 6.82 29.5948 7.798e-08 \*\*\*

## Residuals 593 136.72 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Surprisingly, the adj R2 has increased slightly again. But there is still variable which is insignificant. In this case, **thtr\_rel\_month**. This variable will be removed accordingly.

eighth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+mpaa\_rating+thtr\_rel\_year+audience\_score+

imdb\_num\_votes+critics\_rating+audience\_rating, data=model)

**summary**(eighth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + mpaa\_rating +

## thtr\_rel\_year + audience\_score + imdb\_num\_votes + critics\_rating +

## audience\_rating, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.57778 -0.17827 0.03774 0.25439 1.10313

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.880e+00 4.277e+00 0.907 0.364720

## title\_typeFeature Film -3.480e-01 1.934e-01 -1.799 0.072520 .

## title\_typeTV Movie -3.532e-01 3.094e-01 -1.141 0.254162

## genreAnimation -5.237e-01 1.988e-01 -2.635 0.008642 \*\*

## genreArt House & International 3.405e-01 1.581e-01 2.153 0.031686 \*

## genreComedy -1.256e-01 8.162e-02 -1.539 0.124307

## genreDocumentary 1.266e-01 2.047e-01 0.618 0.536530

## genreDrama 1.527e-01 7.173e-02 2.128 0.033712 \*

## genreHorror 1.175e-01 1.233e-01 0.953 0.341083

## genreMusical & Performing Arts 6.673e-02 1.686e-01 0.396 0.692404

## genreMystery & Suspense 2.950e-01 9.175e-02 3.215 0.001374 \*\*

## genreOther 4.667e-02 1.413e-01 0.330 0.741209

## genreScience Fiction & Fantasy -8.319e-02 1.818e-01 -0.458 0.647422

## runtime 5.019e-03 1.183e-03 4.243 2.56e-05 \*\*\*

## mpaa\_ratingNC-17 1.386e-01 5.033e-01 0.275 0.783112

## mpaa\_ratingPG -1.397e-01 1.423e-01 -0.982 0.326728

## mpaa\_ratingPG-13 -1.677e-01 1.480e-01 -1.133 0.257529

## mpaa\_ratingR -8.890e-02 1.425e-01 -0.624 0.532894

## mpaa\_ratingUnrated -1.532e-01 1.706e-01 -0.898 0.369646

## thtr\_rel\_year -5.043e-05 2.124e-03 -0.024 0.981062

## audience\_score 4.609e-02 2.160e-03 21.341 < 2e-16 \*\*\*

## imdb\_num\_votes 7.503e-07 2.190e-07 3.426 0.000654 \*\*\*

## critics\_ratingFresh -2.859e-02 6.157e-02 -0.464 0.642501

## critics\_ratingRotten -2.979e-01 6.622e-02 -4.499 8.21e-06 \*\*\*

## audience\_ratingUpright -4.287e-01 7.891e-02 -5.433 8.10e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4807 on 594 degrees of freedom

## Multiple R-squared: 0.8077, Adjusted R-squared: 0.8

## F-statistic: 104 on 24 and 594 DF, p-value: < 2.2e-16

**anova**(eighth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 145.8532 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.2735 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.4247 < 2.2e-16 \*\*\*

## mpaa\_rating 5 15.79 3.16 13.6714 1.196e-12 \*\*\*

## thtr\_rel\_year 1 0.98 0.98 4.2545 0.03958 \*

## audience\_score 1 344.63 344.63 1491.6251 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.78 4.78 20.6997 6.517e-06 \*\*\*

## critics\_rating 2 7.20 3.60 15.5747 2.555e-07 \*\*\*

## audience\_rating 1 6.82 6.82 29.5168 8.097e-08 \*\*\*

## Residuals 594 137.24 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As we can see the adj R2 has been hovering around 0.8 and that probably means we are approaching towards the last part of the final regression model. In the final model, we must make sure each and every explanatory variable used in is significant predictors and has low p-value. In my case, using a 95% confidence level, any p-value higher than 0.05 is considered insignificant. Looking at the eighth model above, it seems **thtr\_rel\_year** has very high p-value. I will continue to remove any insignificant predictor in my model to ensure it is a parsimonious model. A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

ninth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+audience\_score+mpaa\_rating+

imdb\_num\_votes+critics\_rating+audience\_rating, data=model)

**summary**(ninth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + audience\_score +

## mpaa\_rating + imdb\_num\_votes + critics\_rating + audience\_rating,

## data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.57790 -0.17857 0.03747 0.25433 1.10365

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.779e+00 2.848e-01 13.265 < 2e-16 \*\*\*

## title\_typeFeature Film -3.479e-01 1.932e-01 -1.800 0.072295 .

## title\_typeTV Movie -3.532e-01 3.092e-01 -1.142 0.253759

## genreAnimation -5.244e-01 1.966e-01 -2.667 0.007866 \*\*

## genreArt House & International 3.405e-01 1.580e-01 2.155 0.031546 \*

## genreComedy -1.256e-01 8.154e-02 -1.540 0.124036

## genreDocumentary 1.265e-01 2.045e-01 0.619 0.536481

## genreDrama 1.526e-01 7.165e-02 2.130 0.033557 \*

## genreHorror 1.177e-01 1.229e-01 0.958 0.338479

## genreMusical & Performing Arts 6.655e-02 1.683e-01 0.395 0.692648

## genreMystery & Suspense 2.950e-01 9.166e-02 3.218 0.001361 \*\*

## genreOther 4.692e-02 1.408e-01 0.333 0.739058

## genreScience Fiction & Fantasy -8.297e-02 1.814e-01 -0.457 0.647588

## runtime 5.025e-03 1.152e-03 4.362 1.52e-05 \*\*\*

## audience\_score 4.609e-02 2.150e-03 21.438 < 2e-16 \*\*\*

## mpaa\_ratingNC-17 1.389e-01 5.027e-01 0.276 0.782385

## mpaa\_ratingPG -1.399e-01 1.420e-01 -0.985 0.325009

## mpaa\_ratingPG-13 -1.684e-01 1.452e-01 -1.159 0.246736

## mpaa\_ratingR -8.941e-02 1.407e-01 -0.635 0.525478

## mpaa\_ratingUnrated -1.541e-01 1.658e-01 -0.930 0.352837

## imdb\_num\_votes 7.491e-07 2.133e-07 3.512 0.000479 \*\*\*

## critics\_ratingFresh -2.832e-02 6.042e-02 -0.469 0.639446

## critics\_ratingRotten -2.978e-01 6.604e-02 -4.510 7.81e-06 \*\*\*

## audience\_ratingUpright -4.287e-01 7.884e-02 -5.438 7.90e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4803 on 595 degrees of freedom

## Multiple R-squared: 0.8077, Adjusted R-squared: 0.8003

## F-statistic: 108.7 on 23 and 595 DF, p-value: < 2.2e-16

**anova**(ninth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.0986 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.3413 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 155.6862 < 2.2e-16 \*\*\*

## audience\_score 1 360.03 360.03 1560.8804 < 2.2e-16 \*\*\*

## mpaa\_rating 5 1.17 0.23 1.0153 0.4077

## imdb\_num\_votes 1 4.95 4.95 21.4775 4.402e-06 \*\*\*

## critics\_rating 2 7.24 3.62 15.6850 2.299e-07 \*\*\*

## audience\_rating 1 6.82 6.82 29.5668 7.896e-08 \*\*\*

## Residuals 595 137.24 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To the last league of the elimination process, **mpaa\_rating** carries an insiginificant p-value. I will attempt to remove this variable and assess the model through adj R2 again.

tenth\_mod <-**lm**(imdb\_rating~title\_type+genre+runtime+audience\_score+imdb\_num\_votes+

critics\_rating+audience\_rating, data=model)

**summary**(tenth\_mod)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54779 -0.18685 0.04423 0.25968 1.05127

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.671e+00 2.544e-01 14.429 < 2e-16 \*\*\*

## title\_typeFeature Film -3.267e-01 1.910e-01 -1.710 0.087747 .

## title\_typeTV Movie -3.411e-01 3.080e-01 -1.107 0.268631

## genreAnimation -4.681e-01 1.818e-01 -2.574 0.010287 \*

## genreArt House & International 3.302e-01 1.541e-01 2.143 0.032484 \*

## genreComedy -1.394e-01 8.050e-02 -1.732 0.083746 .

## genreDocumentary 1.152e-01 2.022e-01 0.569 0.569232

## genreDrama 1.532e-01 6.973e-02 2.197 0.028411 \*

## genreHorror 1.333e-01 1.202e-01 1.109 0.267731

## genreMusical & Performing Arts 6.554e-02 1.672e-01 0.392 0.695270

## genreMystery & Suspense 3.081e-01 8.961e-02 3.438 0.000626 \*\*\*

## genreOther 3.427e-02 1.398e-01 0.245 0.806471

## genreScience Fiction & Fantasy -6.879e-02 1.810e-01 -0.380 0.704109

## runtime 4.728e-03 1.138e-03 4.156 3.71e-05 \*\*\*

## audience\_score 4.627e-02 2.134e-03 21.684 < 2e-16 \*\*\*

## imdb\_num\_votes 7.335e-07 2.103e-07 3.487 0.000523 \*\*\*

## critics\_ratingFresh -3.492e-02 6.009e-02 -0.581 0.561359

## critics\_ratingRotten -3.089e-01 6.536e-02 -4.726 2.86e-06 \*\*\*

## audience\_ratingUpright -4.279e-01 7.855e-02 -5.448 7.46e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4798 on 600 degrees of freedom

## Multiple R-squared: 0.8065, Adjusted R-squared: 0.8007

## F-statistic: 138.9 on 18 and 600 DF, p-value: < 2.2e-16

**anova**(tenth\_mod)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.401 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.425 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 156.008 < 2.2e-16 \*\*\*

## audience\_score 1 360.03 360.03 1564.111 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.79 4.79 20.813 6.143e-06 \*\*\*

## critics\_rating 2 7.69 3.85 16.706 8.695e-08 \*\*\*

## audience\_rating 1 6.83 6.83 29.677 7.456e-08 \*\*\*

## Residuals 600 138.11 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The last model has the highest adj R2 of 0.8007 and all predictors are statistical significant.

Final Model

final\_model <-**lm**(imdb\_rating~title\_type+genre+runtime+audience\_score+imdb\_num\_votes+

critics\_rating+audience\_rating, data=model)

**summary**(final\_model)

##

## Call:

## lm(formula = imdb\_rating ~ title\_type + genre + runtime + audience\_score +

## imdb\_num\_votes + critics\_rating + audience\_rating, data = model)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.54779 -0.18685 0.04423 0.25968 1.05127

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 3.671e+00 2.544e-01 14.429 < 2e-16 \*\*\*

## title\_typeFeature Film -3.267e-01 1.910e-01 -1.710 0.087747 .

## title\_typeTV Movie -3.411e-01 3.080e-01 -1.107 0.268631

## genreAnimation -4.681e-01 1.818e-01 -2.574 0.010287 \*

## genreArt House & International 3.302e-01 1.541e-01 2.143 0.032484 \*

## genreComedy -1.394e-01 8.050e-02 -1.732 0.083746 .

## genreDocumentary 1.152e-01 2.022e-01 0.569 0.569232

## genreDrama 1.532e-01 6.973e-02 2.197 0.028411 \*

## genreHorror 1.333e-01 1.202e-01 1.109 0.267731

## genreMusical & Performing Arts 6.554e-02 1.672e-01 0.392 0.695270

## genreMystery & Suspense 3.081e-01 8.961e-02 3.438 0.000626 \*\*\*

## genreOther 3.427e-02 1.398e-01 0.245 0.806471

## genreScience Fiction & Fantasy -6.879e-02 1.810e-01 -0.380 0.704109

## runtime 4.728e-03 1.138e-03 4.156 3.71e-05 \*\*\*

## audience\_score 4.627e-02 2.134e-03 21.684 < 2e-16 \*\*\*

## imdb\_num\_votes 7.335e-07 2.103e-07 3.487 0.000523 \*\*\*

## critics\_ratingFresh -3.492e-02 6.009e-02 -0.581 0.561359

## critics\_ratingRotten -3.089e-01 6.536e-02 -4.726 2.86e-06 \*\*\*

## audience\_ratingUpright -4.279e-01 7.855e-02 -5.448 7.46e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.4798 on 600 degrees of freedom

## Multiple R-squared: 0.8065, Adjusted R-squared: 0.8007

## F-statistic: 138.9 on 18 and 600 DF, p-value: < 2.2e-16

**anova**(final\_model)

## Analysis of Variance Table

##

## Response: imdb\_rating

## Df Sum Sq Mean Sq F value Pr(>F)

## title\_type 2 67.40 33.70 146.401 < 2.2e-16 \*\*\*

## genre 10 93.05 9.30 40.425 < 2.2e-16 \*\*\*

## runtime 1 35.91 35.91 156.008 < 2.2e-16 \*\*\*

## audience\_score 1 360.03 360.03 1564.111 < 2.2e-16 \*\*\*

## imdb\_num\_votes 1 4.79 4.79 20.813 6.143e-06 \*\*\*

## critics\_rating 2 7.69 3.85 16.706 8.695e-08 \*\*\*

## audience\_rating 1 6.83 6.83 29.677 7.456e-08 \*\*\*

## Residuals 600 138.11 0.23

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

After going through the iterative process above, we have finally reached a model that fits a parsimonious concept. This last model has the highest predictive power (i.e.adj R2 of 0.8007) and all predictors are statitiscally significant(i.e. close to zero p-value). In summary, we can reject our null hypothesis that all the coefficients are indifferent and accept our alternative hypothesis that one of these coefficients are not equal.

Model Diagnostics

Now that we have reached a final model, the next step is to diagnose the model to ensure:  
(i) the residuals are scattered randomly;  
(ii) the residuals are nearly normally distributed;  
(iii) the residuals display constant variability; and  
(iv) the residuals are independent

(i) Residuals scattered randomly

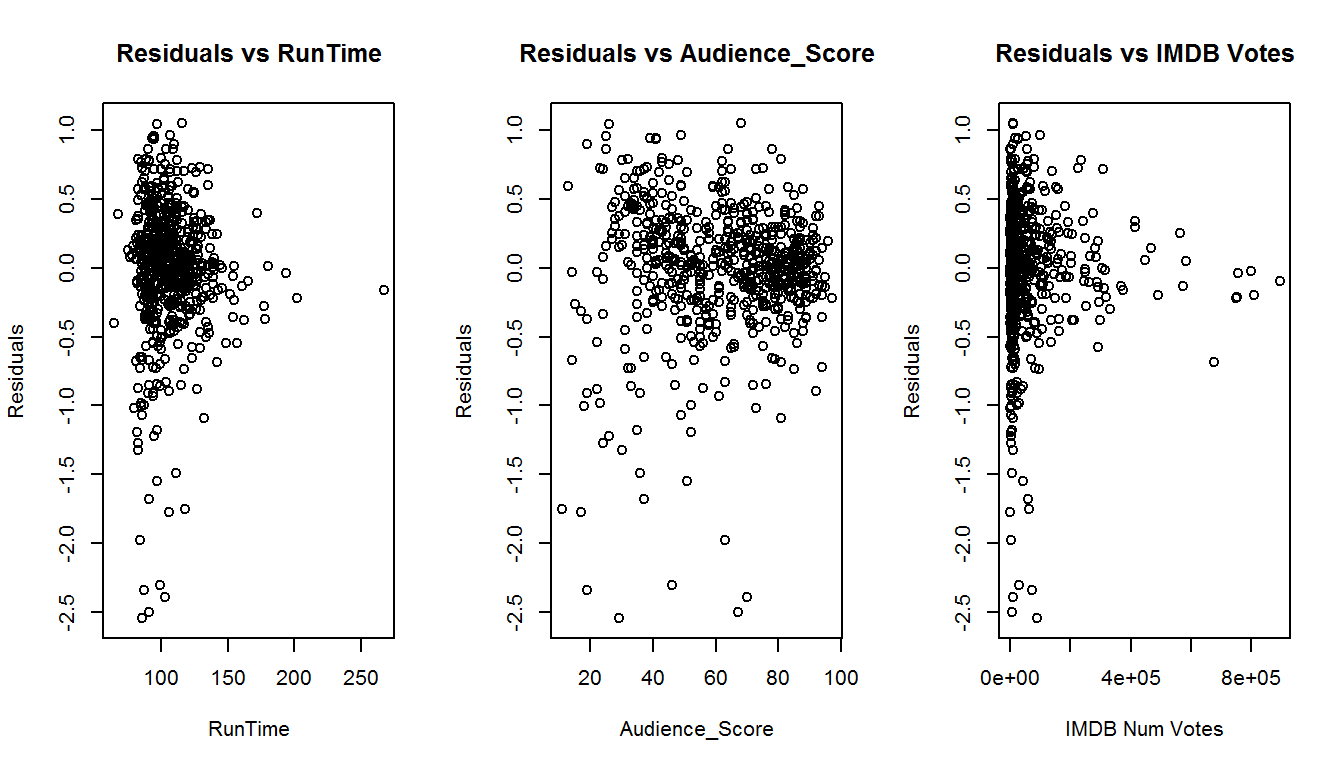
We are check this against all the numerical explanatory variables.

**par**(mfrow = **c**(1, 3))

**plot**(final\_model$residuals~dataset$runtime, ylab="Residuals", xlab="RunTime", main="Residuals vs RunTime")

**plot**(final\_model$residuals~dataset$audience\_score, ylab="Residuals", xlab="Audience\_Score", main="Residuals vs Audience\_Score")

**plot**(final\_model$residuals~dataset$imdb\_num\_votes, ylab="Residuals", xlab="IMDB Num Votes", main="Residuals vs IMDB Votes")



Based on the results above, the residuals scattered randomly with the explanatory variables.

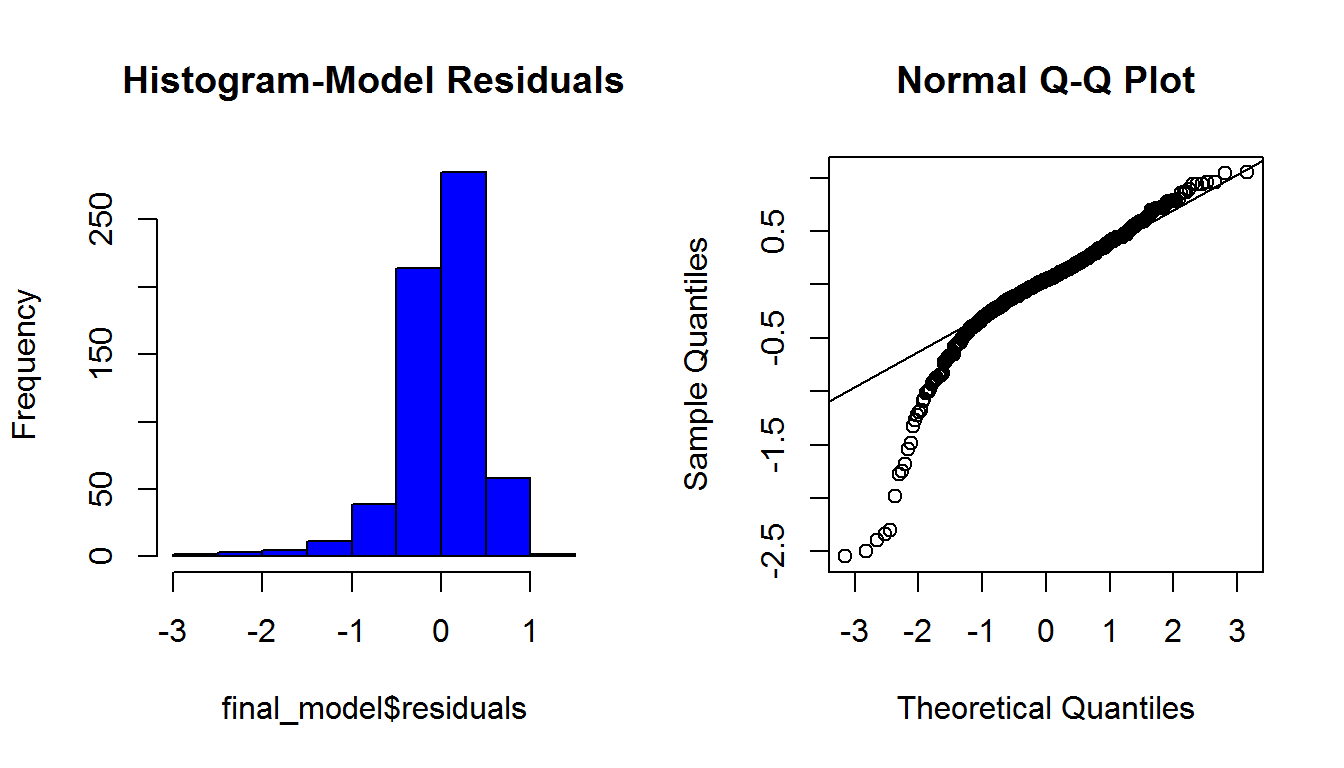
(ii) Residuals are nearly normally distributed

**par**(mfrow = **c**(1, 2))

**hist**(final\_model$residuals, col="blue", main="Histogram-Model Residuals")

**qqnorm**(final\_model$residuals)

**qqline**(final\_model$residuals)



The results of the histogram & qqplot suggest that the residuals are nearly normally distributed around 0 while the qqplot indicates some skewness in the front tail but not significant.

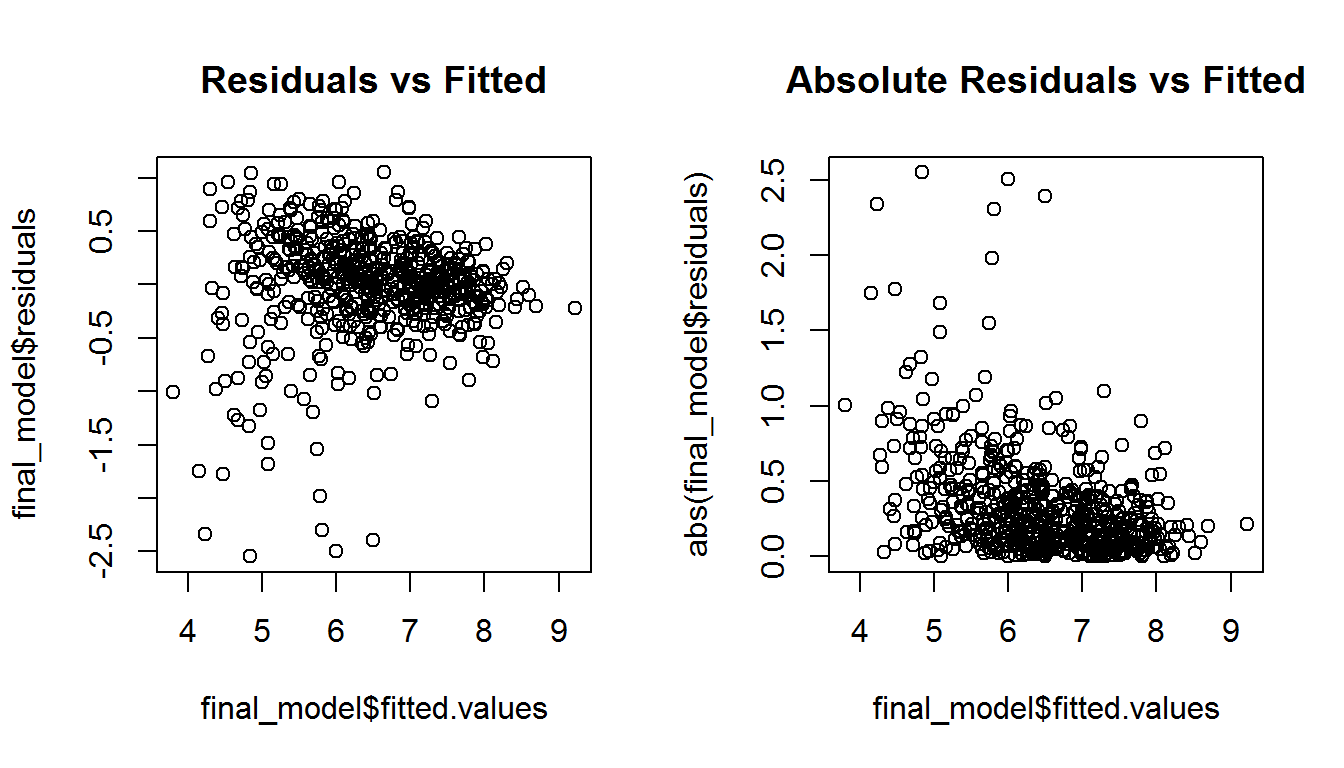
(iii) Residuals display constant variability

This diagnosis is to check homoscedascity and ensure the residuals do not show any structural pattern with the predicted values.

**par**(mfrow = **c**(1, 2))

**plot**(final\_model$residuals~final\_model$fitted.values, main="Residuals vs Fitted")

**plot**(**abs**(final\_model$residuals)~final\_model$fitted.values, main="Absolute Residuals vs Fitted")



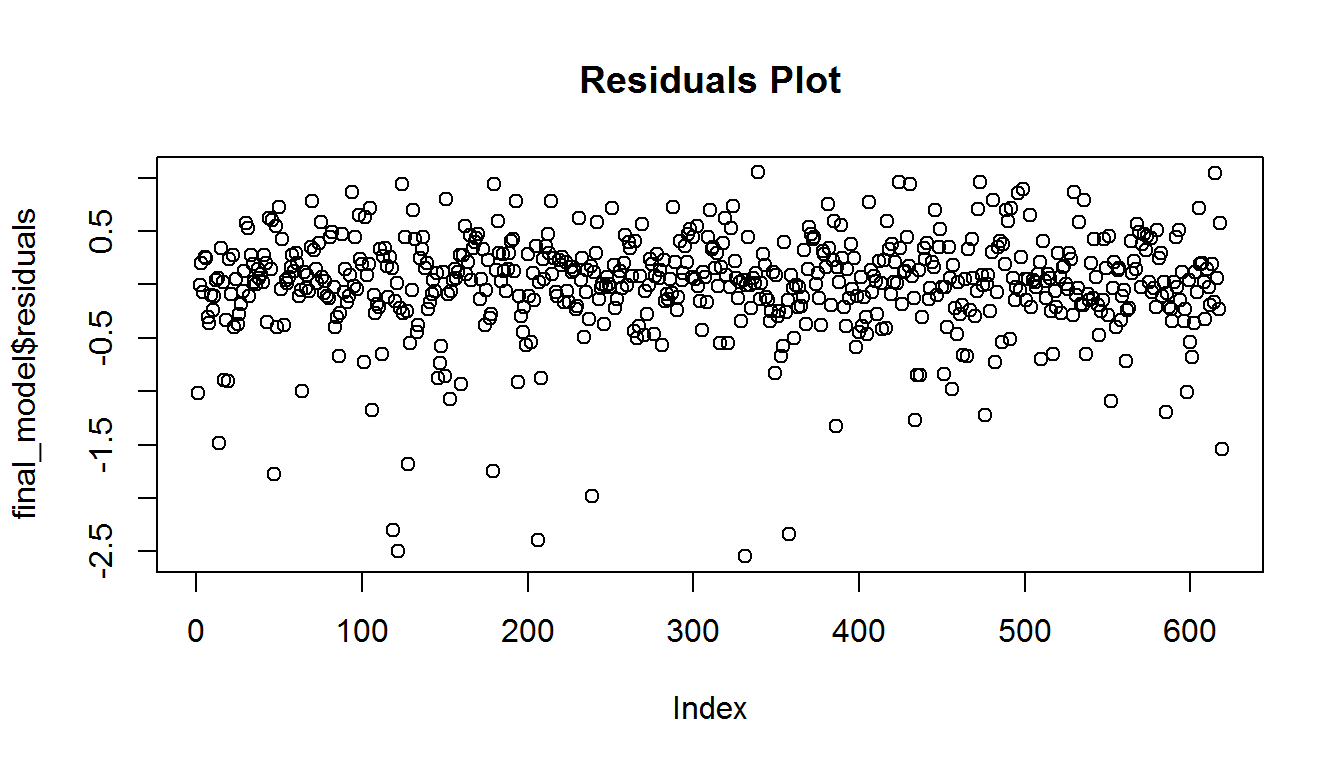
The results show that the residuals do not protray any patterns like fan-shaped or triangle-shaped. It simply means the residuals display constant variability.

(iv) Residuals are independent

The last part of the diagnosis is to plot residuals trended over time. The purpose is mainly to assess if the residuals are independent.

**par**(mfrow = **c**(1, 1))

**plot**(final\_model$residuals, main="Residuals Plot")



Again, the plot showed no specific pattern when trended over time. In summary, this model has satisfied all the assumptions required for a regression model.

Part 5: Prediction

Now that the final model has been developed, this part is to assess its predictive capability. I have selected two movies which are not in the sample but prior to 2016 release (i.e.out-of-sample validation). We will use the model developed to predict the IMDB rating. Two movies are chosen from both [IMDB](https://rstudio-pubs-static.s3.amazonaws.com/www.imdb.com) & [Rotten Tomatoes](https://rstudio-pubs-static.s3.amazonaws.com/www.rottentomatoes.com) websites.

**Movie #1: Armageddon (1998)**  
\* Actual IMDb Rating: 6.6  
\* <http://www.imdb.com/title/tt0120591/?ref_=nv_sr_2>  
\* <https://www.rottentomatoes.com/m/armageddon>

**Movie #2: The Smurfs (2011)**  
\* Actual IMDb Rating: 5.5  
\* <http://www.imdb.com/title/tt0472181/?ref_=nv_sr_3>  
\* <https://www.rottentomatoes.com/m/the_smurfs>

| **Characteristics** | **Armageddon** | **The Smurfs** |
| --- | --- | --- |
| Thtr\_rel\_year | 1998 | 2011 |
| Title\_type | Feature Film | Feature Film |
| Genre | Action & Adventure | Animation |
| Runtime | 151 | 103 |
| Audience Score | 73 | 44 |
| IMDB Num Votes | 335,680 | 69,572 |
| Critics Rating | Rotten | Rotten |
| Audience Rating | Upright | Spilled |

**Movie #1: Armageddon (1998)**

df\_Arma <- **data.frame**(title\_type="Feature Film",genre="Action & Adventure",runtime=151,audience\_score=73,imdb\_num\_votes=335680,critics\_rating="Rotten",audience\_rating="Upright")

**predict**(final\_model, df\_Arma, interval="prediction")

## fit lwr upr

## 1 6.945032 5.985225 7.904839

The predicted value is 6.9 which is very close to the actual IMDb rating of 6.6. In fact, with 95% confidence interval, the actual IMDb\_rating for **Armageddon** has a lower bound of 5.98 and a higher bound of 7.9.

**Movie #2: The Smurfs (2011)**

df\_Smurfs <- **data.frame**(title\_type="Feature Film",genre="Animation",runtime=103,audience\_score=44,imdb\_num\_votes=69572,critics\_rating="Rotten",audience\_rating="Spilled")

**predict**(final\_model, df\_Smurfs, interval="prediction")

## fit lwr upr

## 1 5.140877 4.138902 6.142853

The predicted value is 5.1 which is very close to the actual IMDb rating of 5.5. In fact, with 95% confidence interval, the actual IMDb\_rating for **The Smurfs** has a lower bound of 4.14 and a higher bound of 6.14.

The model has very strong prediction power for the two movies above.The predicted values are so close to the actual IMDB ratings. In fact, both actual ratings lied between the prediction intervals, under 95% confidence level.

Part 6: Conclusion

In conclusion, the model developed herein has a very strong predictive power and can be used to predict the IMDb rating for a particular movie. Going back to my research question, there are strong association on certain specific characteristics of a movie to drive success.This can provide an interesting insights to the producers before they decide what movie to be launched and the likelihood of its success in public eyes.

However, there are some shortcomings to this model:

1. Sample Size & Sampling Method - Given sufficient time, we should discuss about the sample size and how it can represent the entire population. Also, discussion around the sampling method should be addressed so that it can imply causality.
2. Treating Leverage vs Influential Outliers - In my analysis, this is probably something I have discussed very little upon. However, this can be an important part of the information collected as it can distort the overall model accuracy and prediction power.
3. Timing / Information Availability - In my analysis, nothing was mentioned about how soon the data information is readily available in Rotten Tomatoes vs IMDb websites. This model is highly dependent upon these information and it is strongly believed that Rotten Tomatoes information may **only** be available around the same timing as IMDb. In that case, some of these predictor variables may not be available then for future prediction.

Upon addressing the shortcomings of this analysis, there is likelihood that the model developed can be improved further. But as of now, **voila!** we made through our first born….