Linear and Regression Model of Movies Data

Setting working directory and load the movies data set

```
setwd("C:/Users/ddddd/Regression/_movies")
```

```
moviesdata <- load("C:/Users/ddddd/Regression/_movies/_movies.R")
```

Installing packages and loading the libraries

```
library(devtools)
devtools::install_github("statswithr/statsr")
```

```
## Installation failed: Could not resolve host: raw.githubusercontent.com
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(statsr)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)
```

About the Data

Dataset has the collection from IMDB and Rotten Tomatoes websites. Here we are asked to analyse this data so we can find that what thing makes a movie popular? whether it’s name , it’s cast, it’s director, it’s genre ,i.e. whether it is a comedy movie, fictional, romantic, horror, etc.

Both websites are working for the same thing but having different methods.

Rotten Tomatoes is a reviewer website. It gives the reviews based on their critic reviews and popularity in fans and news rarrings etc. It is basically working for the tomatometer rating which breaks it’s review criteria into different zones for example Rotten reviews which have lower ratings , certified fresh reviews have higher rattings and the fresh reviews have the rattings between these two. So by this kind of information one can easily determine and go for a movie or show which is worth watching.

IDMB is also doing the similar things but it also includes some other imformations. It includes the movie rattings, shows and provides the critic reviews and upcoming best movies.

Part 2: Research question

There are various movies which are more popular than others so we are asked to find out this thing. So, I would like to do research about Rotten tomatoes website’s working.So the research question is:

Which elements give the most impact on the popularity of a movie on this website?

Here, I am taking few varibles for this research: genre , critics_score, critics_rating, audience_score, audience_rating, best_dir_win, best_actress_win, best_actor_win, mpaa_rating, runtime etc.

Part 3: Exploratory data analysis

Now, we will see the summary and plot the graph of each variables behavior or the clear view of representing the data. For this first I remove the varibles which are not usable to find the research or we can say the varibles should be excluded and for this purpose we first clean thias data set by removing those varibles.

Removing the varibles which are not in use for research purpose

```
moviesdata$title<-NULL
```

```
## Warning in moviesdata$title <- NULL: Coercing LHS to a list
```

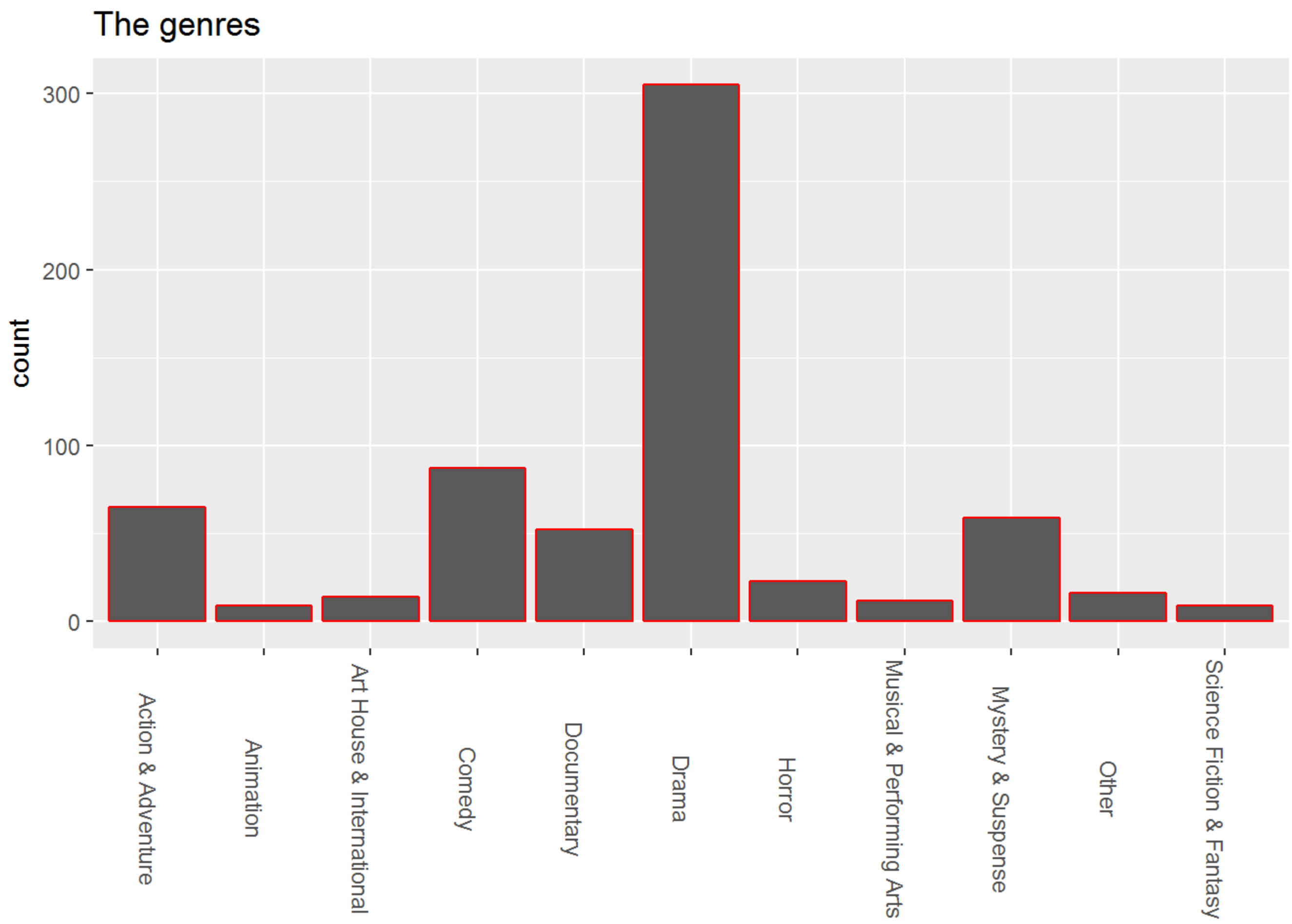
```
moviesdata$best_pic_nom<-NULL
moviesdata$thtr_rel_year<-NULL
moviesdata$title_type<-NULL
moviesdata$studioimdb_rating<-NULL
moviesdata$dvd_rel_year<-NULL
moviesdata$dvd_rel_month<-NULL
moviesdata$dvd_rel_day<-NULL
moviesdata$imdb_num_votes<-NULL
moviesdata$best_pic_win<-NULL
moviesdata$top200_box<-NULL
moviesdata$actor1<-NULL
moviesdata$actor2<-NULL
moviesdata$actor3<-NULL
moviesdata$actor4<-NULL
moviesdata$actor5<-NULL
moviesdata$imdb_url<-NULL
moviesdata$rt_url<-NULL
moviesdata$director<-NULL
moviesdata$dvd_rel_year<-NULL
moviesdata$dvd_rel_month<-NULL
moviesdata$dvd_rel_day<-NULL
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.rotten
tomatoes.com/m/waiting_for_guffman/" "//www.rottentomatoes.com/m/age_of_innocence/" ...
```

Now start plotting the graphs.

1.Plot of genres of movies

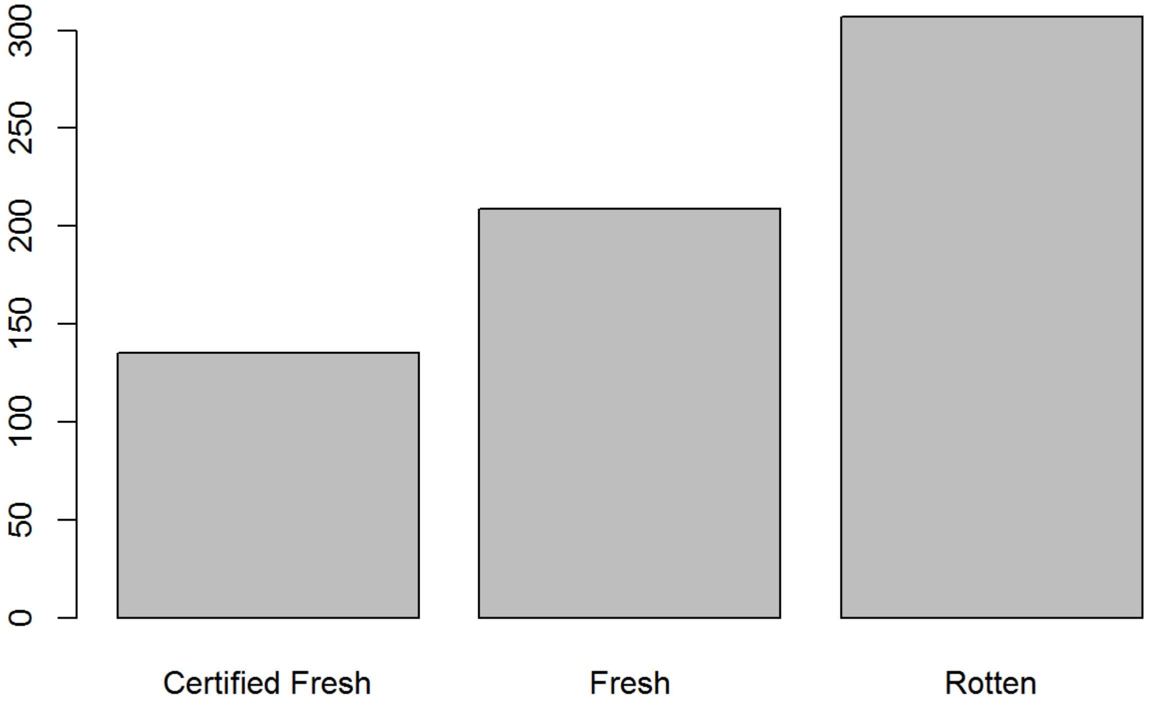
```
library(ggplot2)
ggplot(data=movies , aes(x=genre) ) + geom_bar(color = "red") + ggtitle("The genres") + theme(axis.title.x=element_blank())
 + theme(axis.text.x = element_text(angle=270))
```



By this plot we observe the highest genre is Drama.

Plot of ratings of critics

```
plot(movies$critics_rating)
```

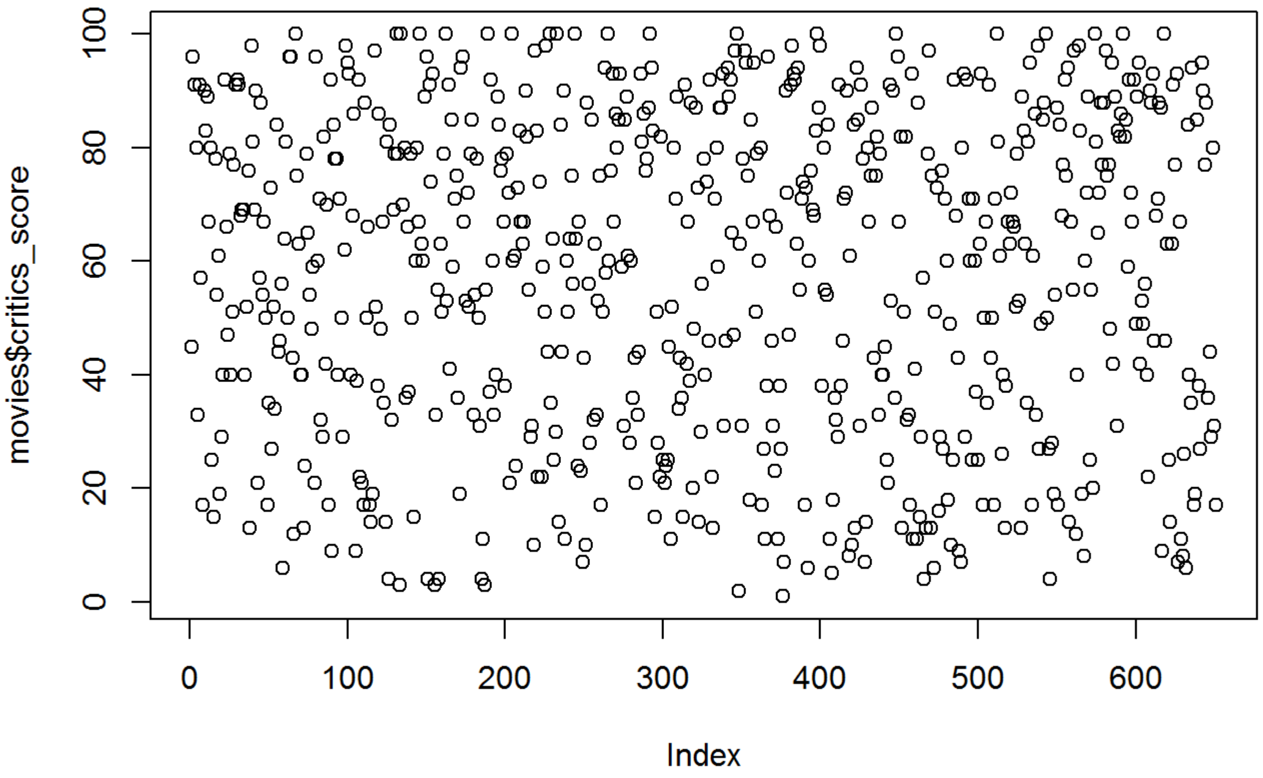


This plot tells the three types ratings

of rotten tomatoes website.

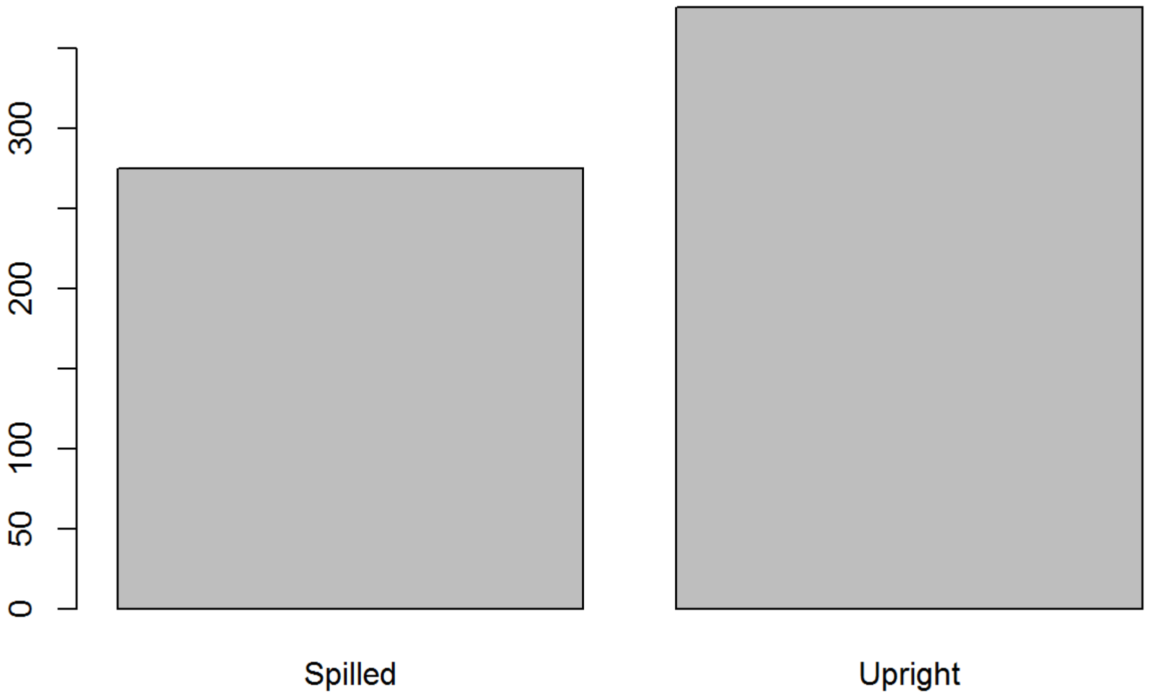
Plot of score of critics

```
plot(movies$critics_score)
```



Plot of ratings of audience

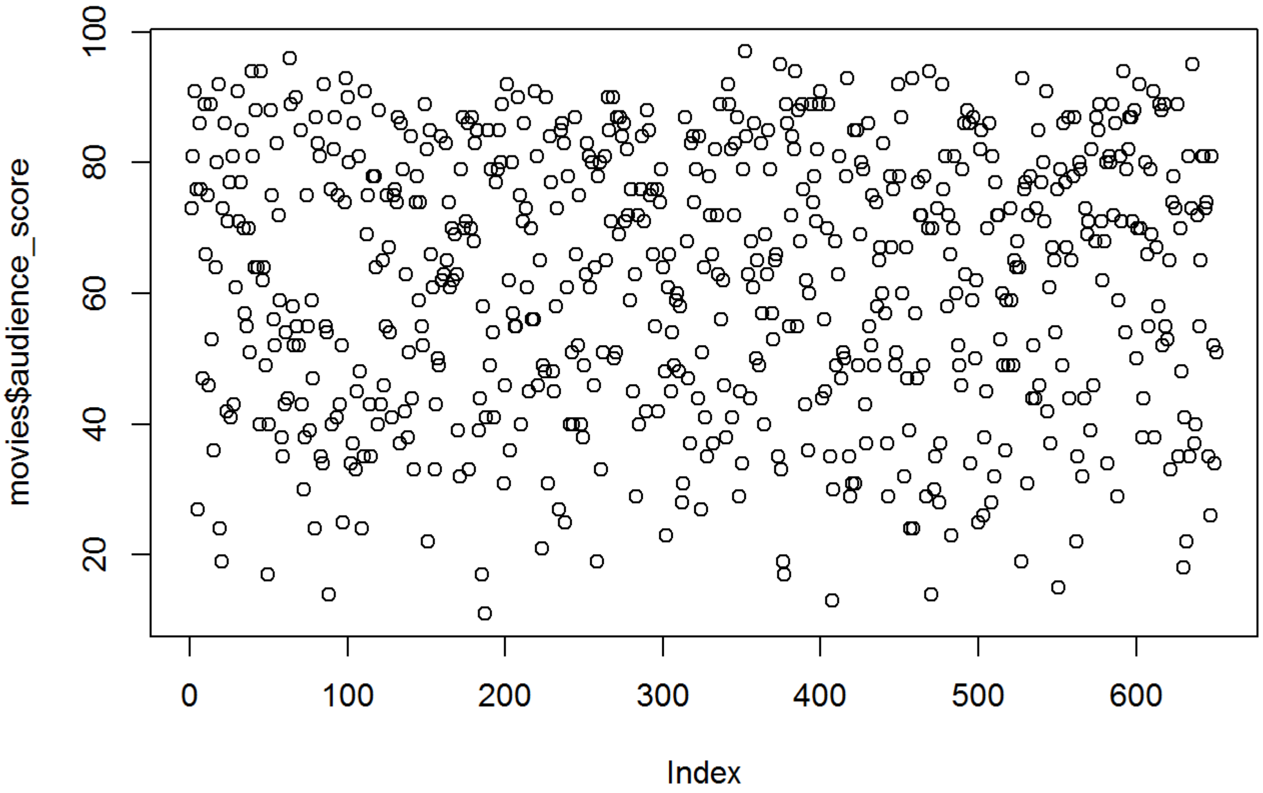
```
plot(movies$audience_rating)
```



This graph tells the audience ratings which is either upright or spilled.

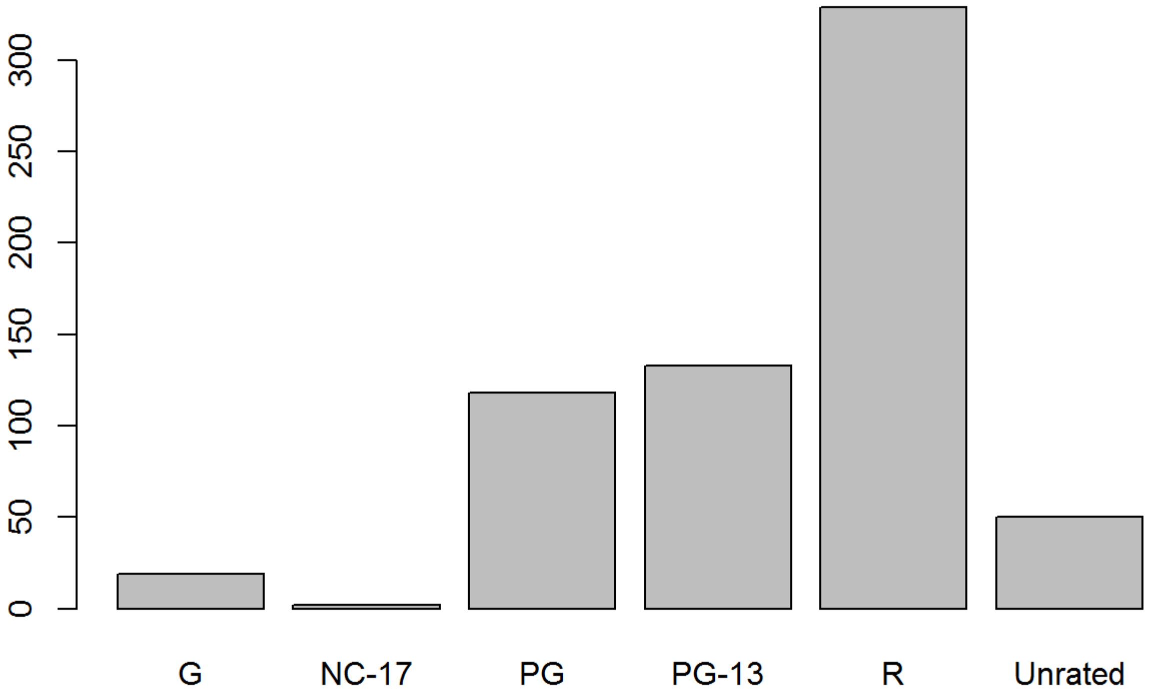
Plot of score of audience

```
plot(movies$audience_score)
```



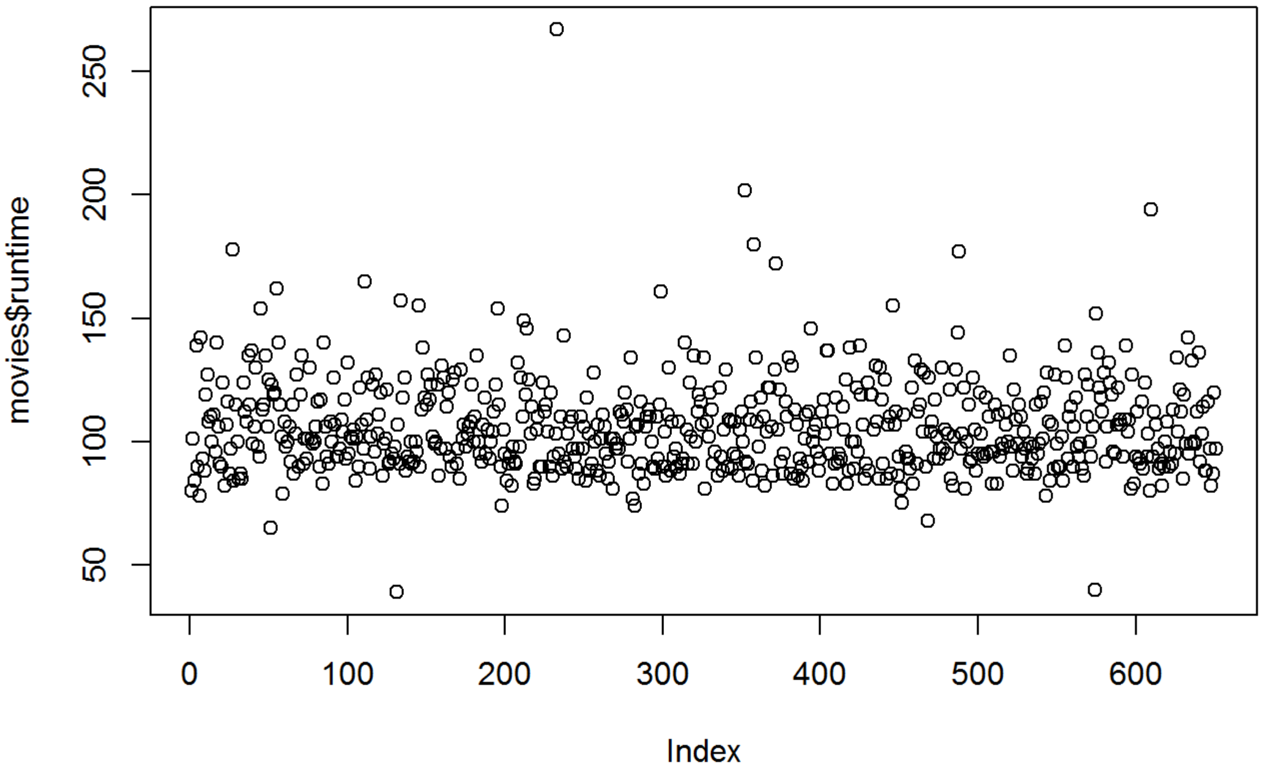
Plot of ratings of MPAA

```
plot(movies$mpaa_rating)
```



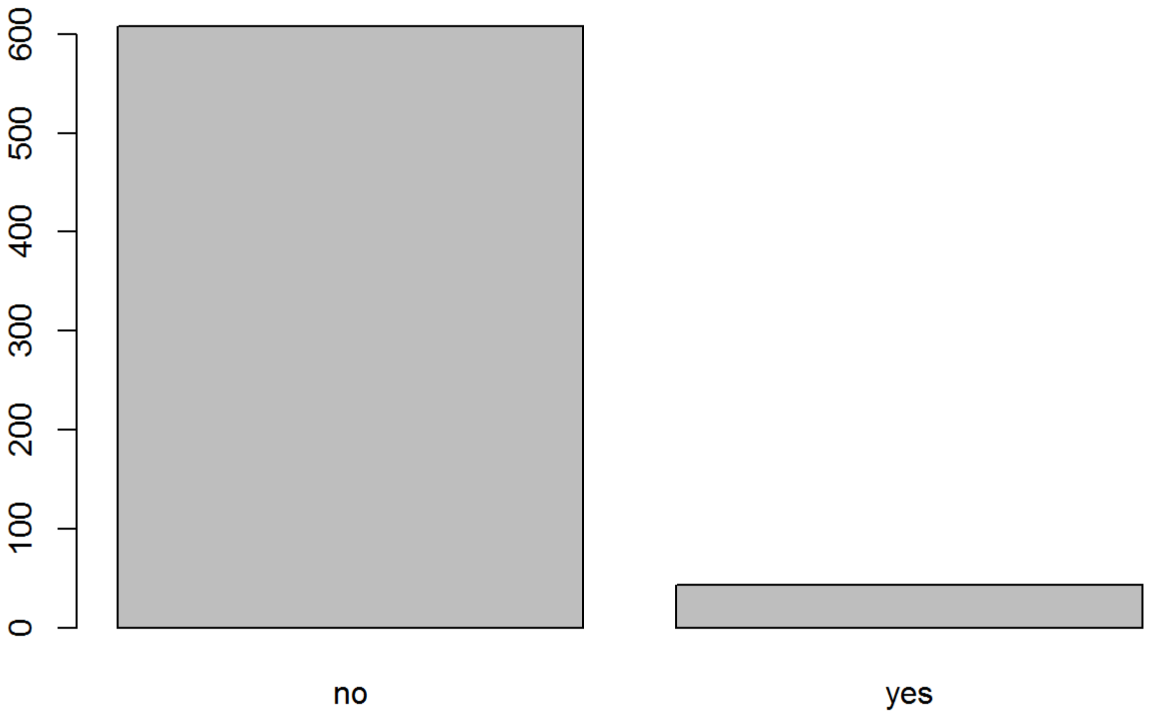
Plot of runtime of movie

```
plot(movies$runtime)
```



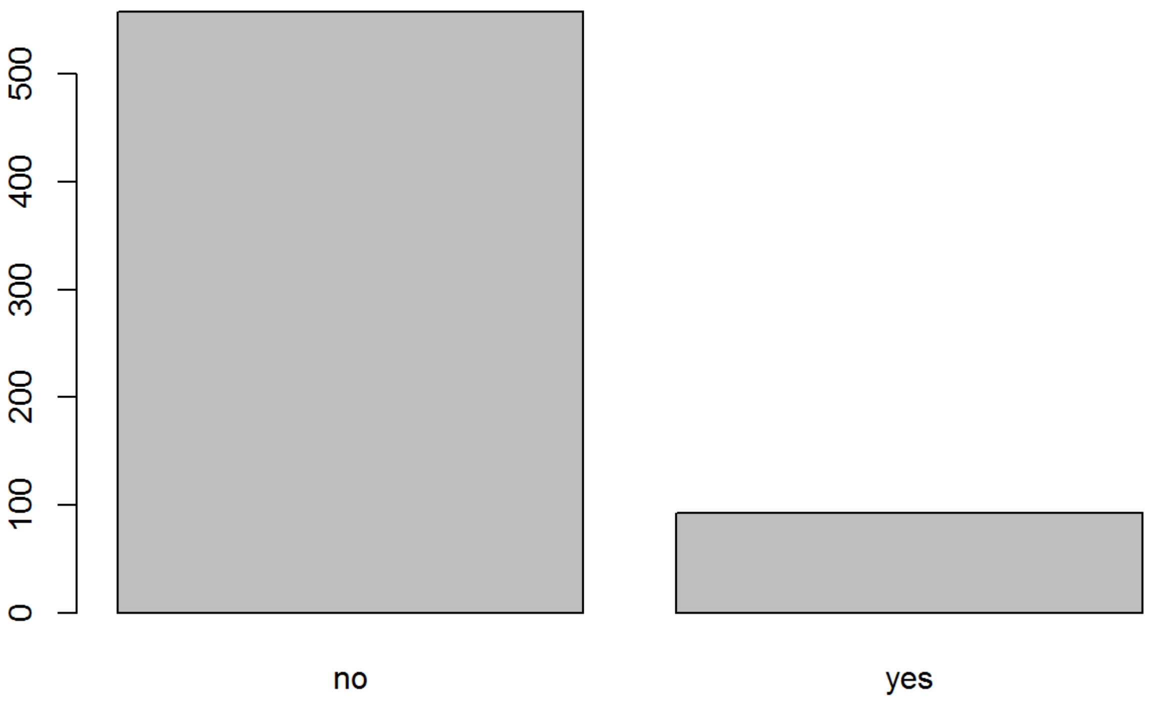
Plot of best director win the award

```
plot(movies$best_dir_win)
```



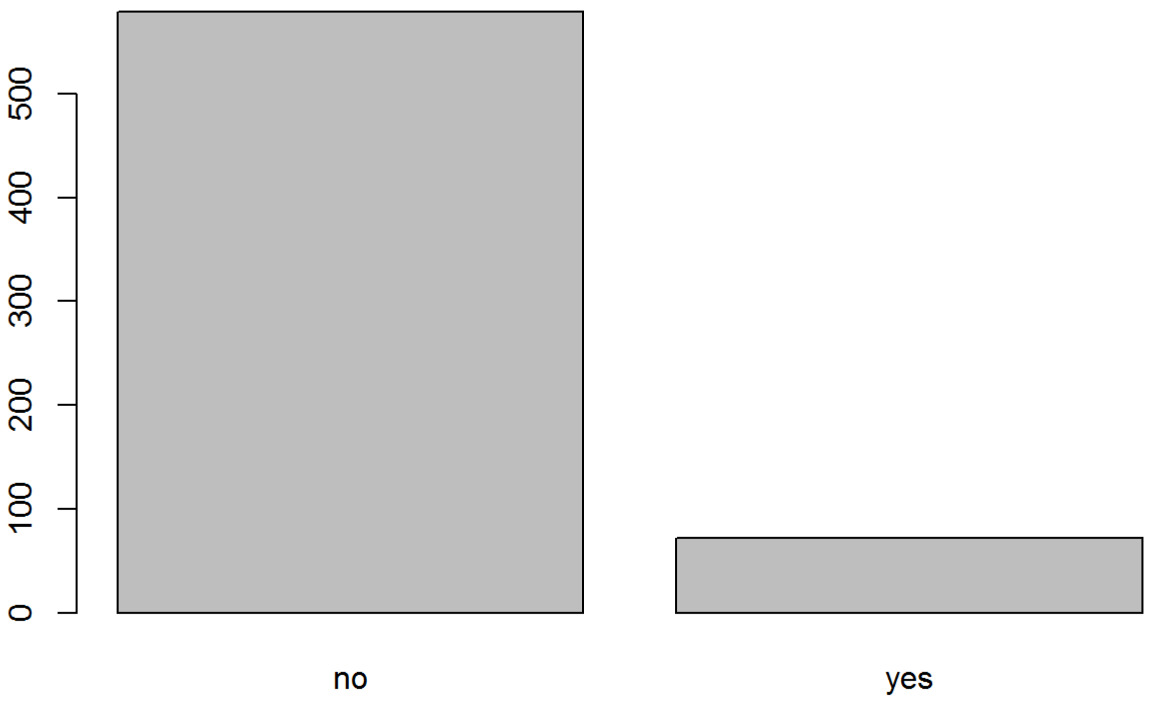
Plot of best actor win the award

```
plot(movies$best_actor_win)
```



Plot of best actress win the award

```
plot(movies$best_actress_win)
```



After this EDA(Exploratory data analysis) , we see some facts of varibles used but the thing which impact the most ,i.e. for a movie rating which varibles are most essential , can’t explained . so move on to the regression model which gives the best results to be predicted.

Part 4 Modelling

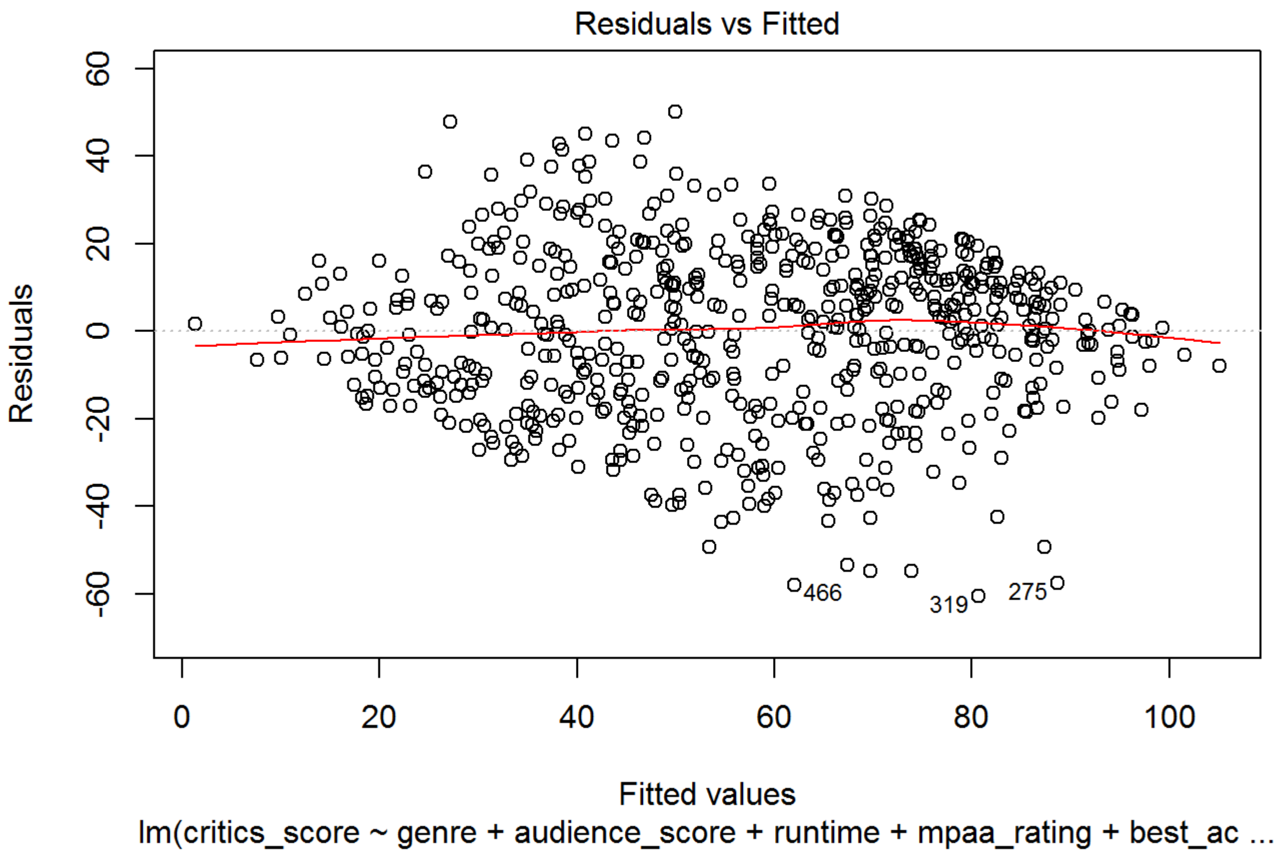
Making a linear regression model, using lm() function. The main purpose is to find what makes a movie popular and critics score and , audience audience score effected it most. It mean while judging a movie to be worth watching we should see these varibles and this will give a best review. Here, I am applying multiple regression modelling in which every further step will exclude one varible and final model give the best prediction.

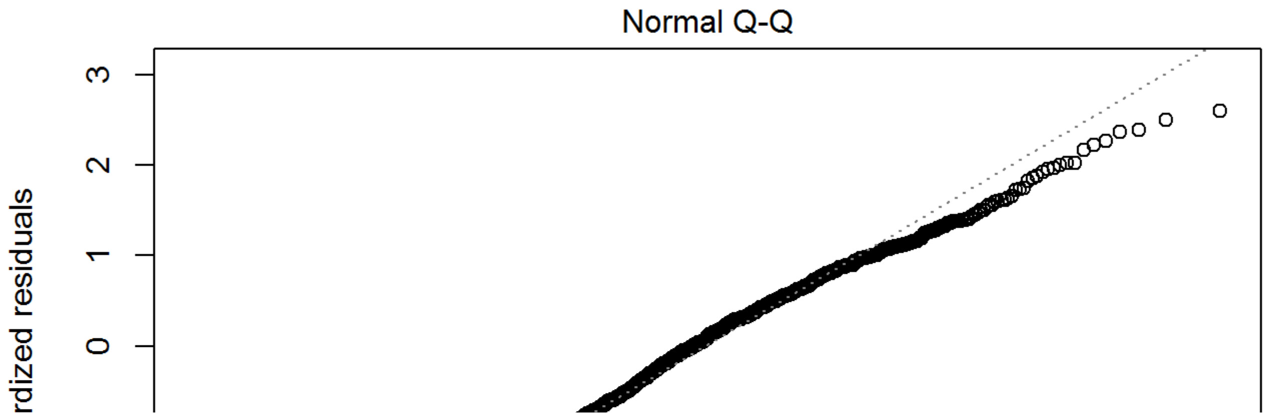
First we include all the variables we see above for checking Note : Although critics rating , audience rating is also giving a clear view to predict a movie but here we don’t take both of them because they are factor and we include only numerical varibles.

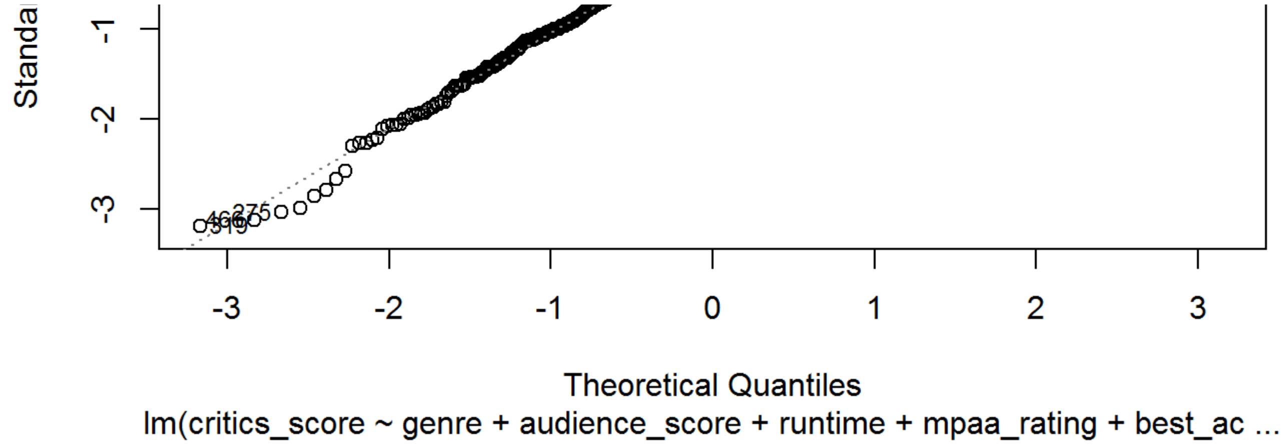
```
my_model_1 = lm( critics_score ~ genre + audience_score + runtime + mpaa_rating + best_actor_win + best_actress_win + best_d
ir_win  , data = movies)
summary(my_model_1)
```

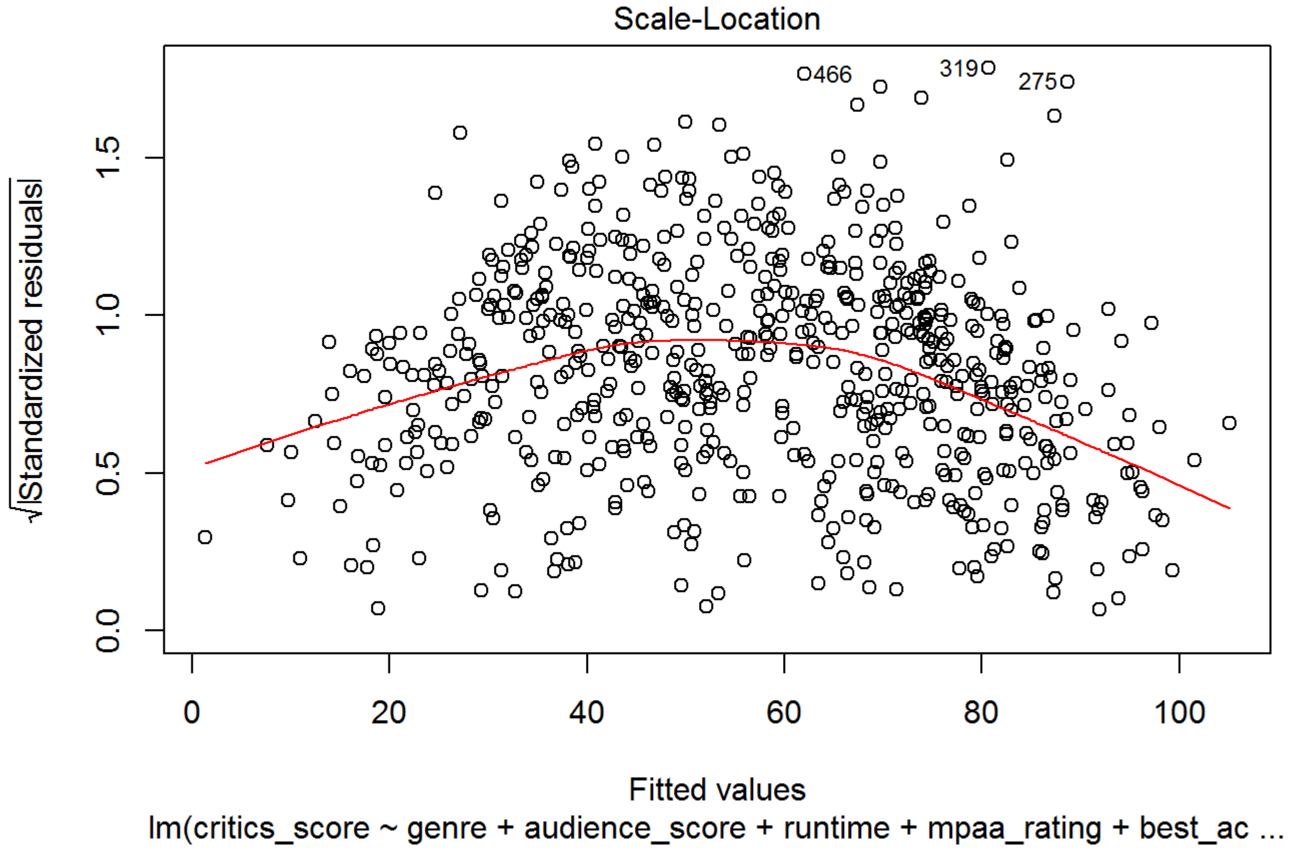
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score + runtime +
##     mpaa_rating + best_actor_win + best_actress_win + best_dir_win,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.672 -13.125   1.842  14.101  50.053
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.04525    7.04947  -0.006 0.994880
## genreAnimation                 -4.78674    7.57186  -0.632 0.527503
## genreArt House & International  -0.64172    5.86868  -0.109 0.912962
## genreComedy                     1.90202    3.23782   0.587 0.557121
## genreDocumentary               14.52471    4.38894   3.309 0.000988 ***
## genreDrama                     10.13705    2.79319   3.629 0.000307 ***
## genreHorror                     9.21518    4.84200   1.903 0.057474 .
## genreMusical & Performing Arts  10.42039    6.23919   1.670 0.095387 .
## genreMystery & Suspense         10.92235    3.61634   3.020 0.002628 **
## genreOther                     10.97696    5.47413   2.005 0.045365 *
## genreScience Fiction & Fantasy   9.88282    6.90969   1.430 0.153132
## audience_score                  0.85884    0.04324  19.864  < 2e-16 ***
## runtime                         0.04423    0.04570   0.968 0.333486
## mpaa_ratingNC-17               13.41119   14.71416   0.911 0.362409
## mpaa_ratingPG                  -8.60914    5.34083  -1.612 0.107475
## mpaa_ratingPG-13              -14.37010    5.47551  -2.624 0.008890 **
## mpaa_ratingR                   -9.97495    5.29058  -1.885 0.059834 .
## mpaa_ratingUnrated             -0.05602    6.06020  -0.009 0.992628
## best_actor_winyes               1.07550    2.29570   0.468 0.639599
## best_actress_winyes             2.85205    2.52700   1.129 0.259484
## best_dir_winyes                 8.00555    3.18927   2.510 0.012318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.38 on 629 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.5346
## F-statistic: 38.28 on 20 and 629 DF,  p-value: < 2.2e-16
```

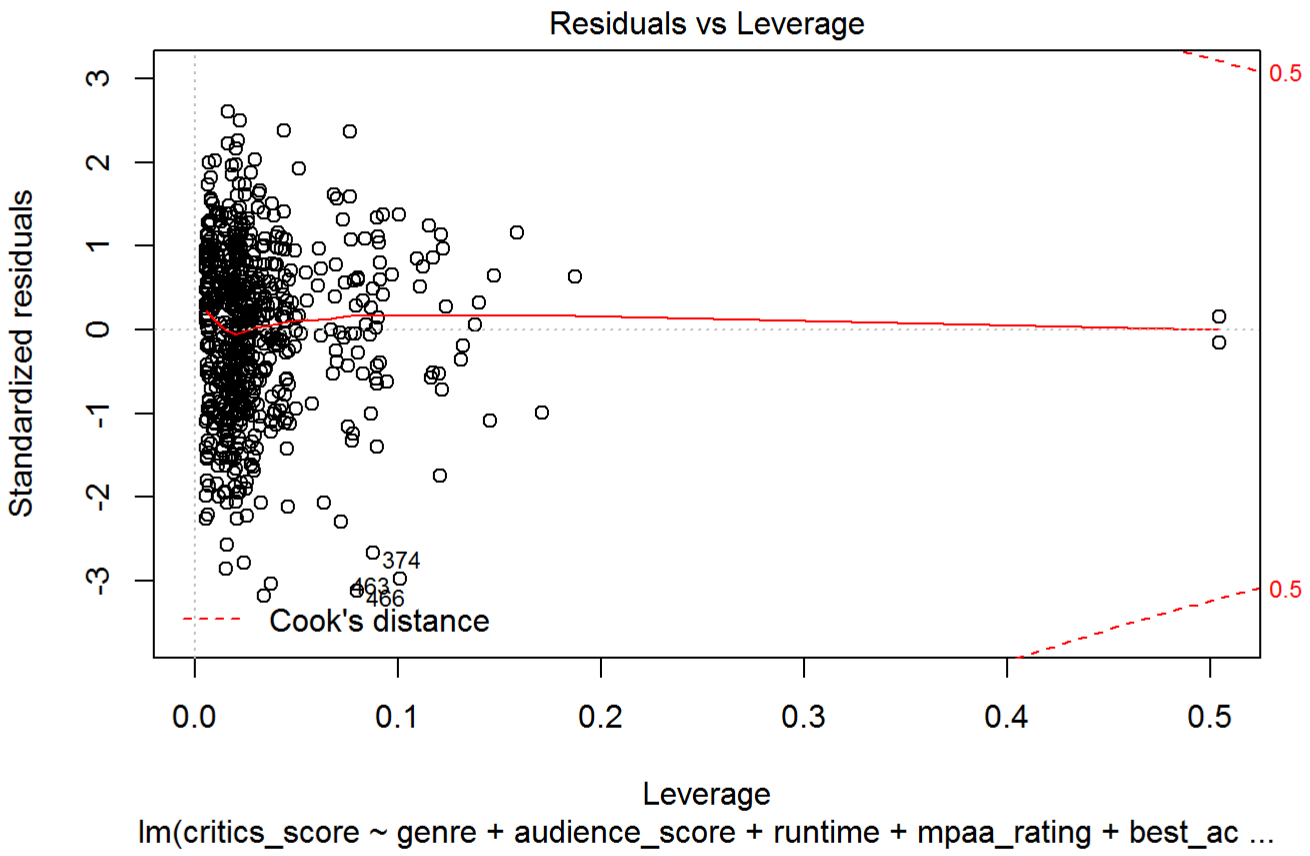
```
plot(my_model_1)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
466
319
275
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + runtime + mpaa_rating + best_ac ...


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
466
319
275
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score + runtime + mpaa_rating + best_ac ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
466
319
275
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + runtime + mpaa_rating + best_ac ...

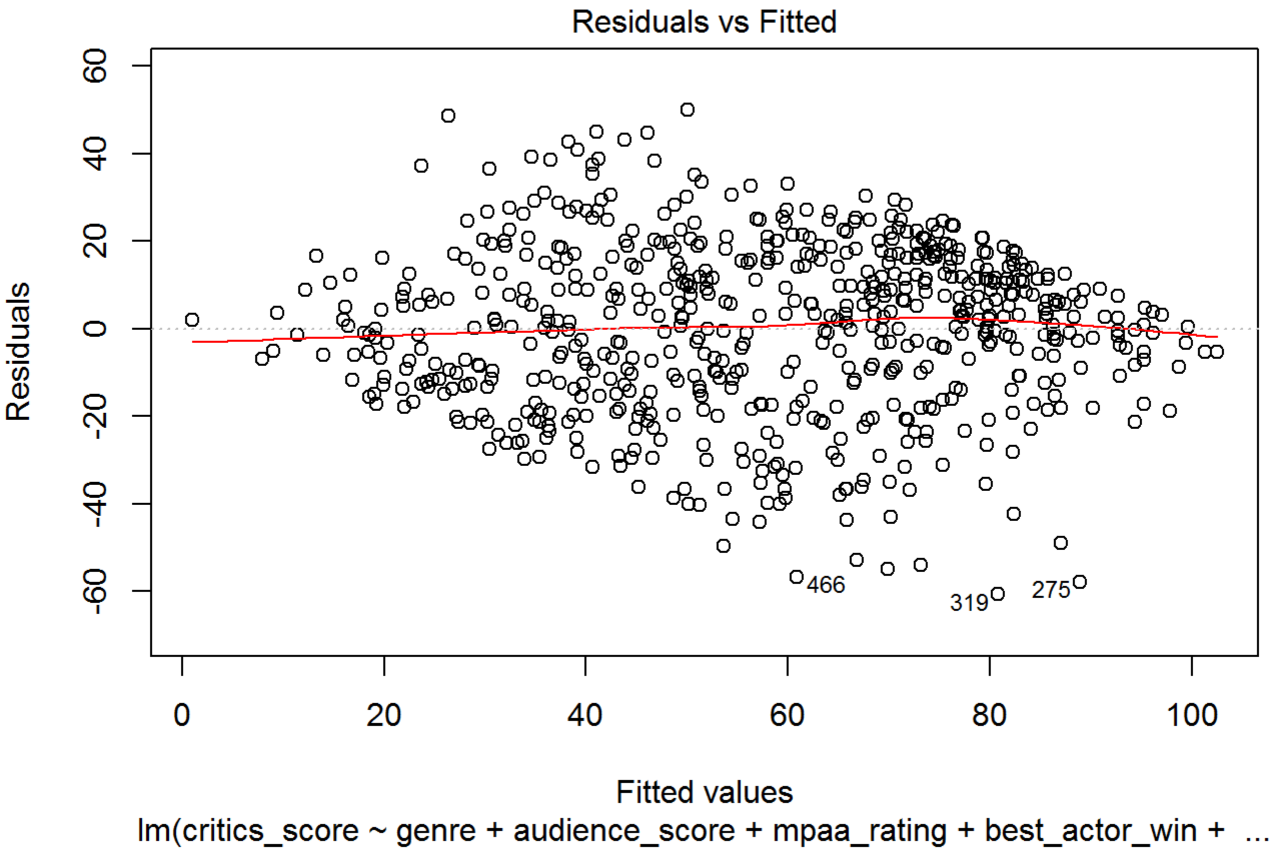


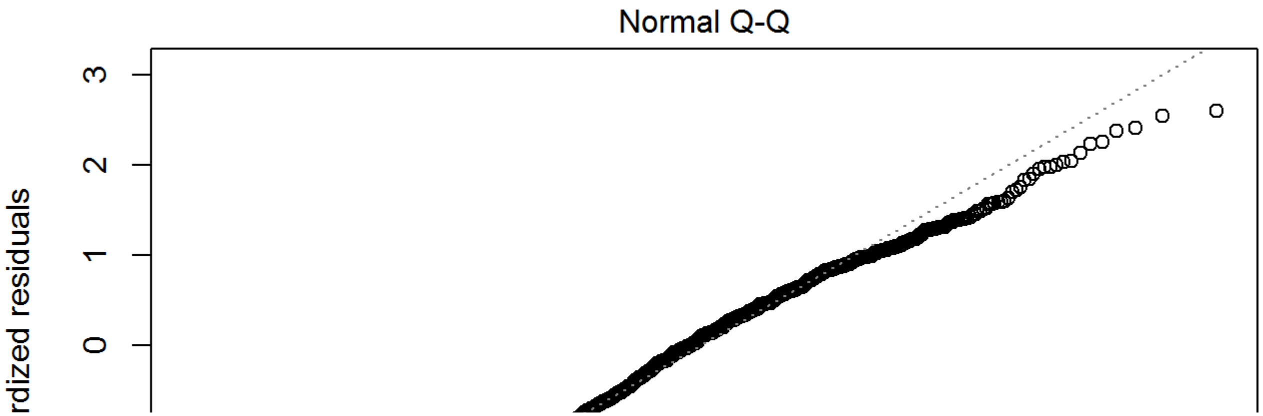
Excluding runtime

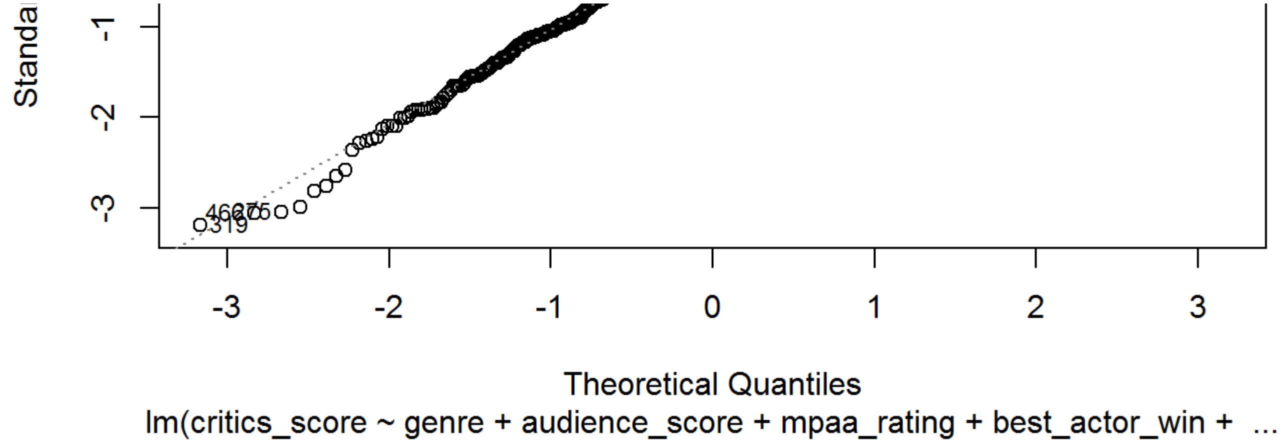
```
my_model_2 = lm( critics_score ~ genre + audience_score + mpaa_rating + best_actor_win + best_actress_win + best_dir_win  ,
data = movies)
summary(my_model_2)
```

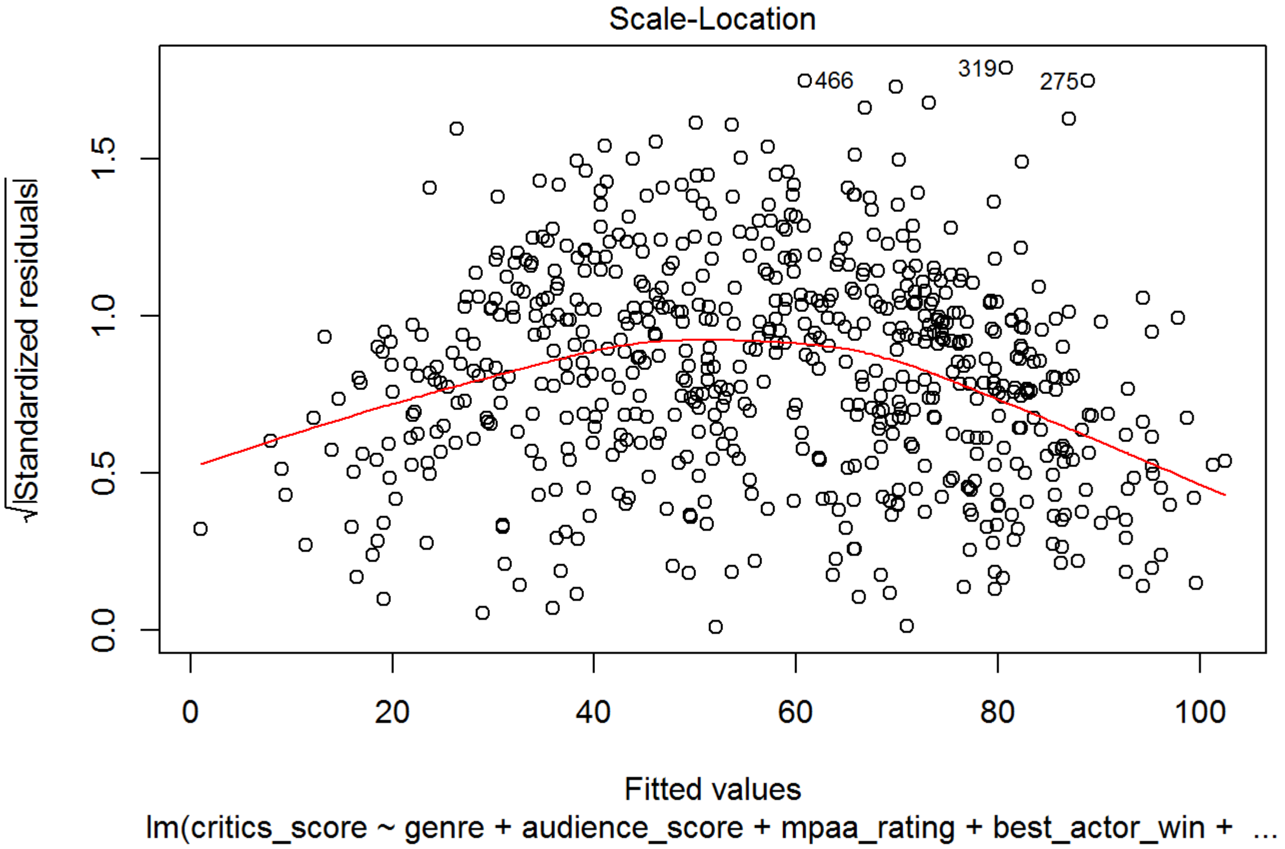
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score + mpaa_rating +
##     best_actor_win + best_actress_win + best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.802 -12.907   2.065  14.030  49.978
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.60588    5.94952   0.606  0.54468
## genreAnimation                  -5.23492    7.55131  -0.693  0.48841
## genreArt House & International  -0.79714    5.86105  -0.136  0.89186
## genreComedy                      1.52622    3.21173   0.475  0.63481
## genreDocumentary                13.95637    4.33970   3.216  0.00137 **
## genreDrama                      10.21927    2.78955   3.663  0.00027 ***
## genreHorror                      8.80000    4.81863   1.826  0.06829 .
## genreMusical & Performing Arts  10.63920    6.22983   1.708  0.08817 .
## genreMystery & Suspense         10.99697    3.61247   3.044  0.00243 **
## genreOther                      11.12750    5.46731   2.035  0.04224 *
## genreScience Fiction & Fantasy   9.85156    6.90379   1.427  0.15408
## audience_score                   0.86586    0.04258  20.336  < 2e-16 ***
## mpaa_ratingNC-17                13.42780   14.70169   0.913  0.36141
## mpaa_ratingPG                   -8.22866    5.32172  -1.546  0.12255
## mpaa_ratingPG-13               -13.64626    5.41941  -2.518  0.01205 *
## mpaa_ratingR                    -9.49213    5.26231  -1.804  0.07174 .
## mpaa_ratingUnrated               0.61910    6.00596   0.103  0.91793
## best_actor_winyes                1.53031    2.24522   0.682  0.49575
## best_actress_winyes              3.20563    2.49835   1.283  0.19993
## best_dir_winyes                  8.62149    3.12234   2.761  0.00593 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.37 on 631 degrees of freedom
## Multiple R-squared:  0.5487, Adjusted R-squared:  0.5351
## F-statistic: 40.38 on 19 and 631 DF,  p-value: < 2.2e-16
```

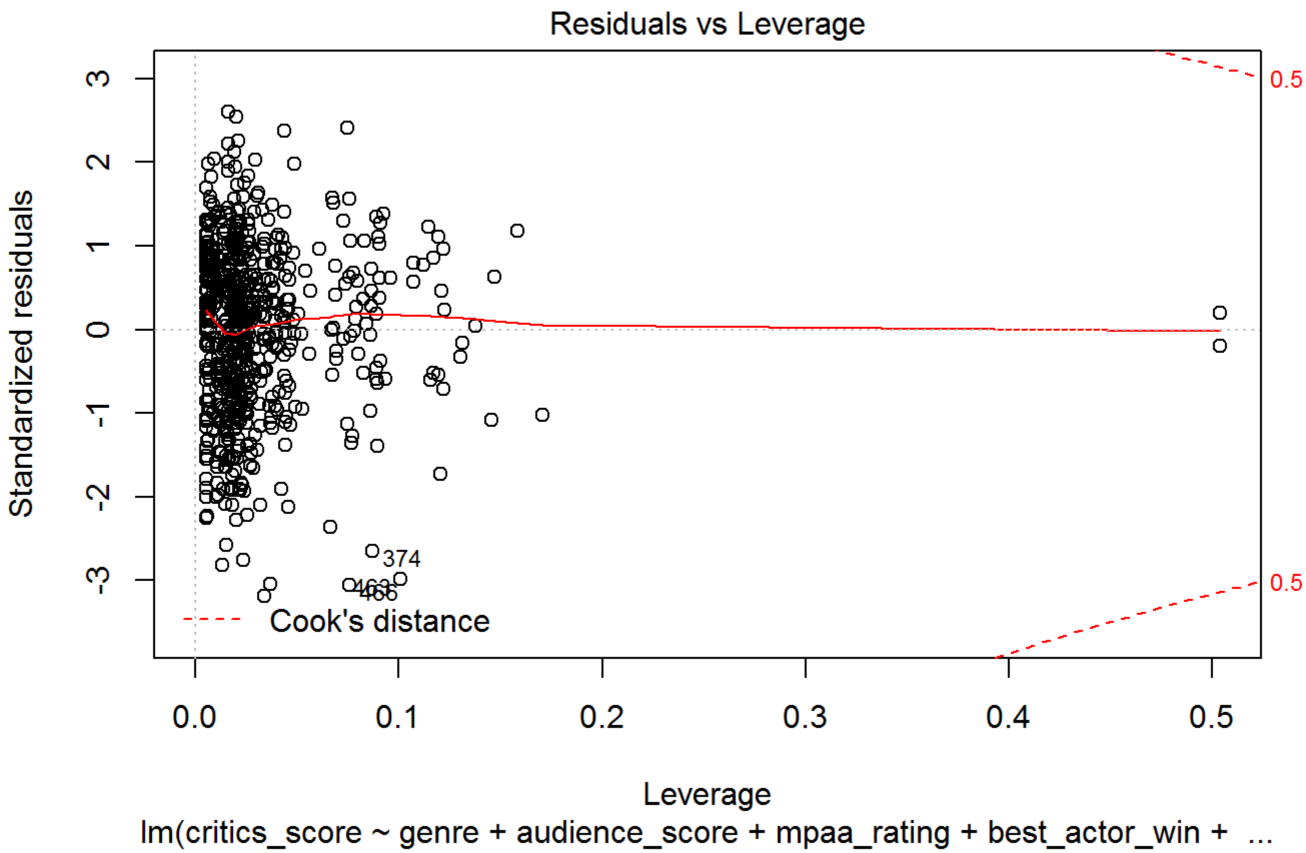
```
plot(my_model_2)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
466
319
275
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + mpaa_rating + best_actor_win + ...


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
466
275
319
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score + mpaa_rating + best_actor_win + ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
466
319
275
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + mpaa_rating + best_actor_win + ...

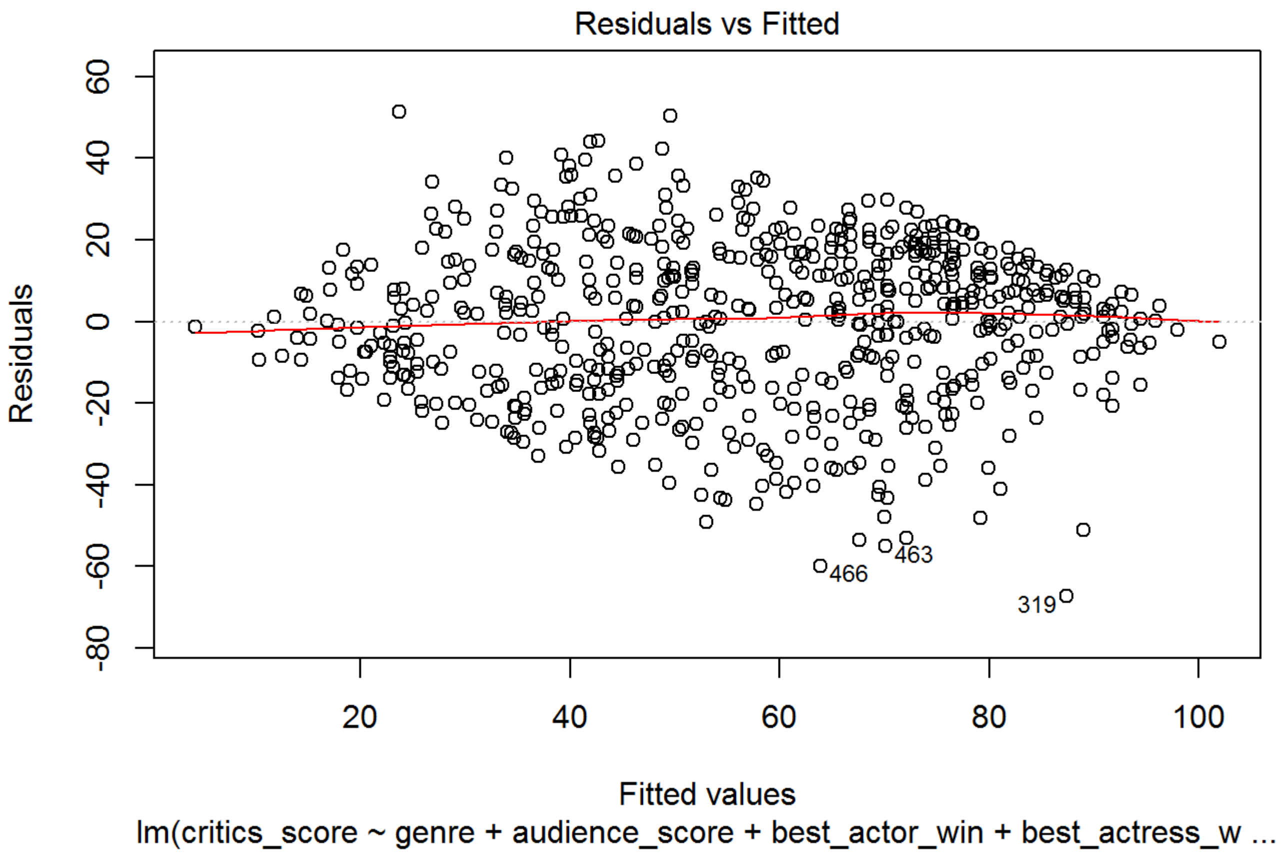


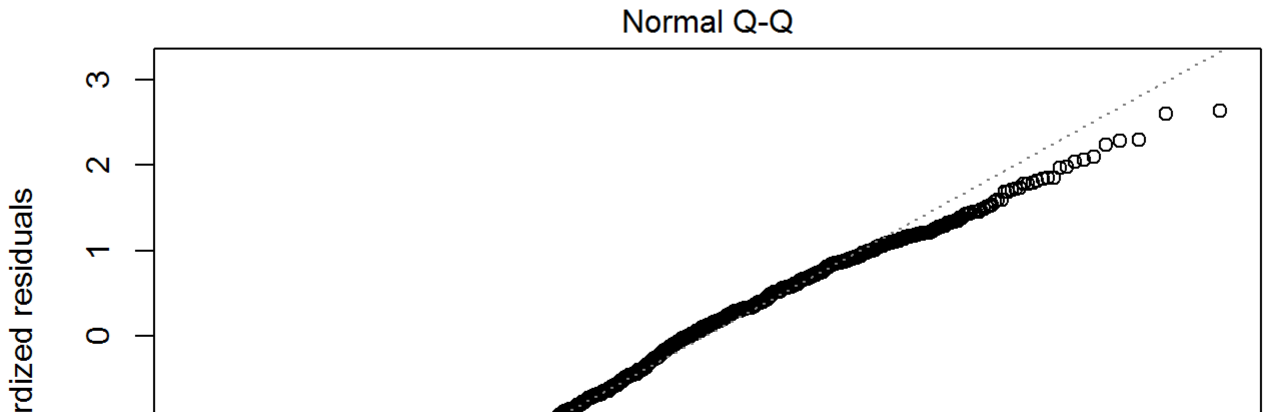
Excluding mpaa_rating

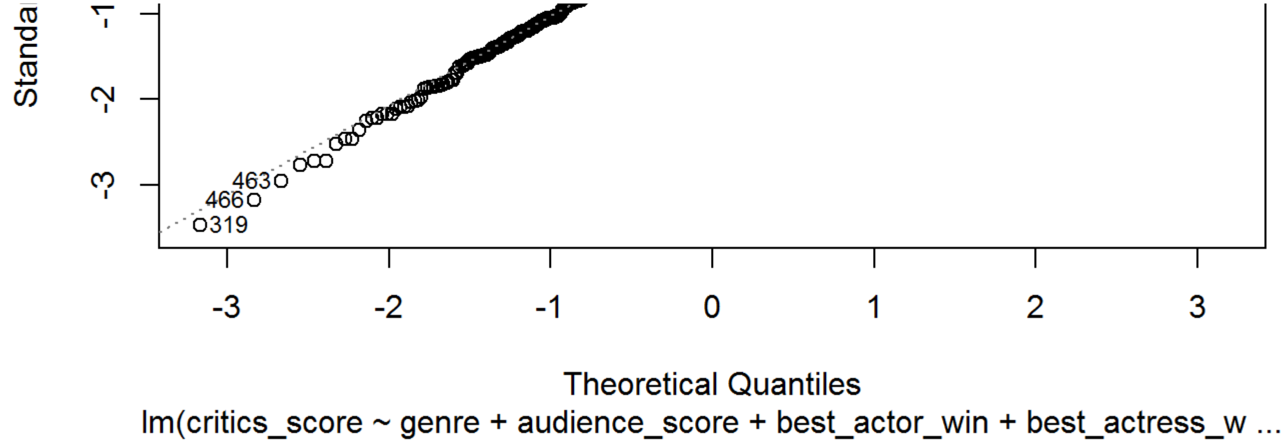
```
my_model_2 = lm( critics_score ~ genre + audience_score  + best_actor_win + best_actress_win + best_dir_win  , data = movie
s)
summary(my_model_2)
```

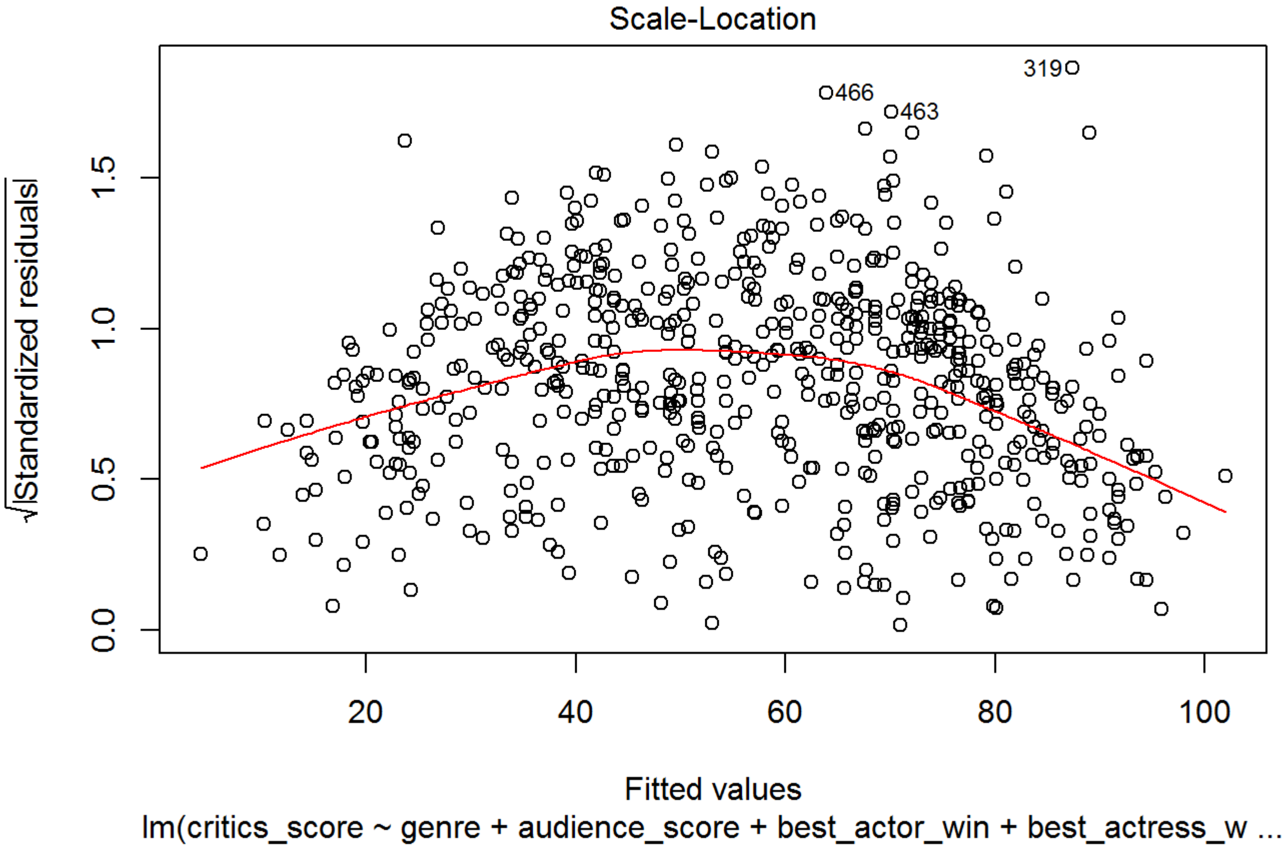
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score + best_actor_win +
##     best_actress_win + best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.349 -13.205   2.321  13.996  51.309
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -7.00748    3.33661  -2.100 0.036106 *
## genreAnimation                  1.28884    6.98535   0.185 0.853675
## genreArt House & International  1.59645    5.80043   0.275 0.783230
## genreComedy                     0.49226    3.22303   0.153 0.878658
## genreDocumentary               19.72801    3.86318   5.107 4.34e-07 ***
## genreDrama                      9.81710    2.74913   3.571 0.000382 ***
## genreHorror                     9.89863    4.76954   2.075 0.038352 *
## genreMusical & Performing Arts 11.65608    6.25490   1.864 0.062851 .
## genreMystery & Suspense        10.55507    3.56317   2.962 0.003168 **
## genreOther                     11.51734    5.50373   2.093 0.036777 *
## genreScience Fiction & Fantasy 10.89645    6.97362   1.563 0.118662
## audience_score                  0.88843    0.04263  20.838  < 2e-16 ***
## best_actor_winyes               1.44884    2.26060   0.641 0.521812
## best_actress_winyes             2.72136    2.51958   1.080 0.280514
## best_dir_winyes                 8.09840    3.15193   2.569 0.010416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.59 on 636 degrees of freedom
## Multiple R-squared:  0.5348, Adjusted R-squared:  0.5245
## F-statistic: 52.22 on 14 and 636 DF,  p-value: < 2.2e-16
```

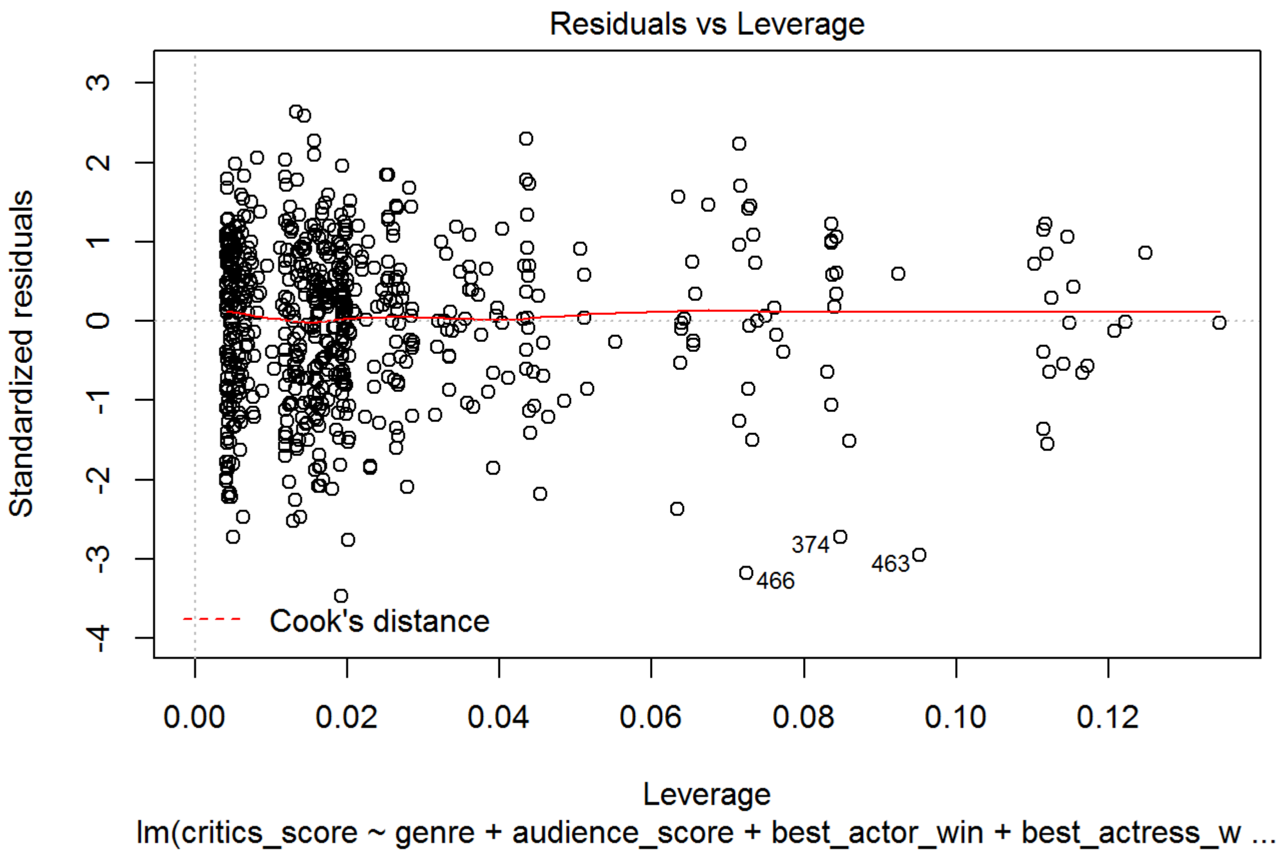
```
plot(my_model_2)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
-80
466
463
319
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + best_actor_win + best_actress_w ...


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
463
466
319
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score + best_actor_win + best_actress_w ...


Scale-Location
319
466
463
0.0
0.5
1.0
1.5
√|Standardized residuals|
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + best_actor_win + best_actress_w ...

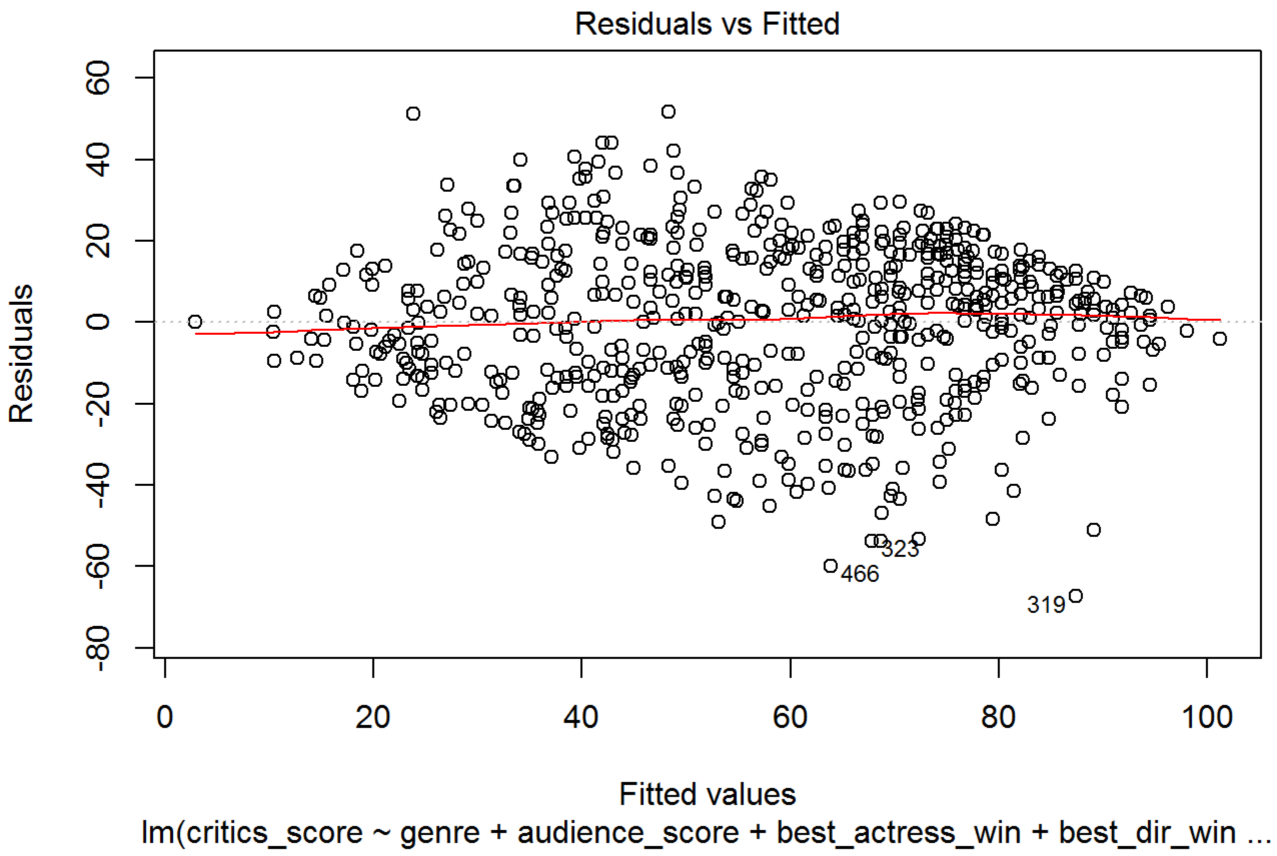


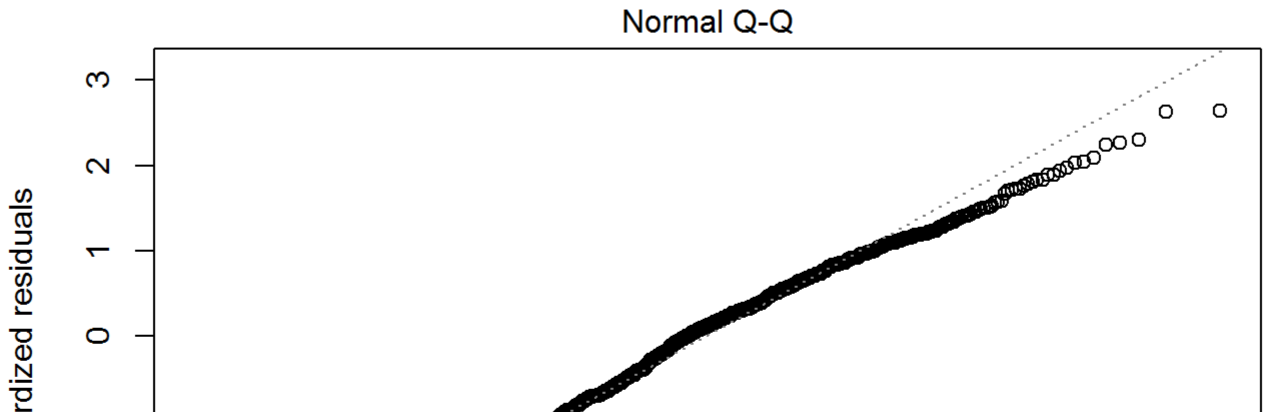
Excluding best_actor_win

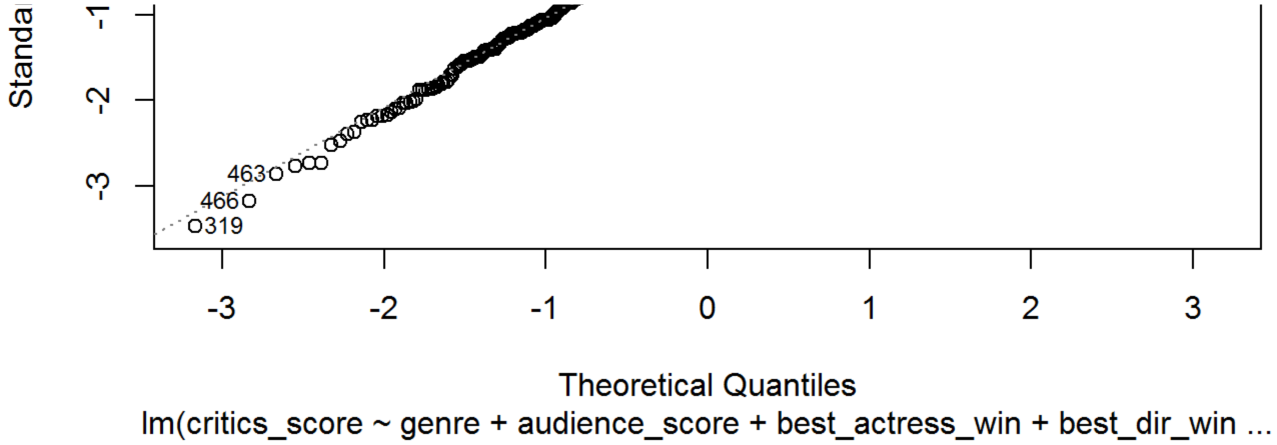
```
my_model_3 = lm( critics_score ~ genre + audience_score  + best_actress_win + best_dir_win  , data = movies)
summary(my_model_3)
```

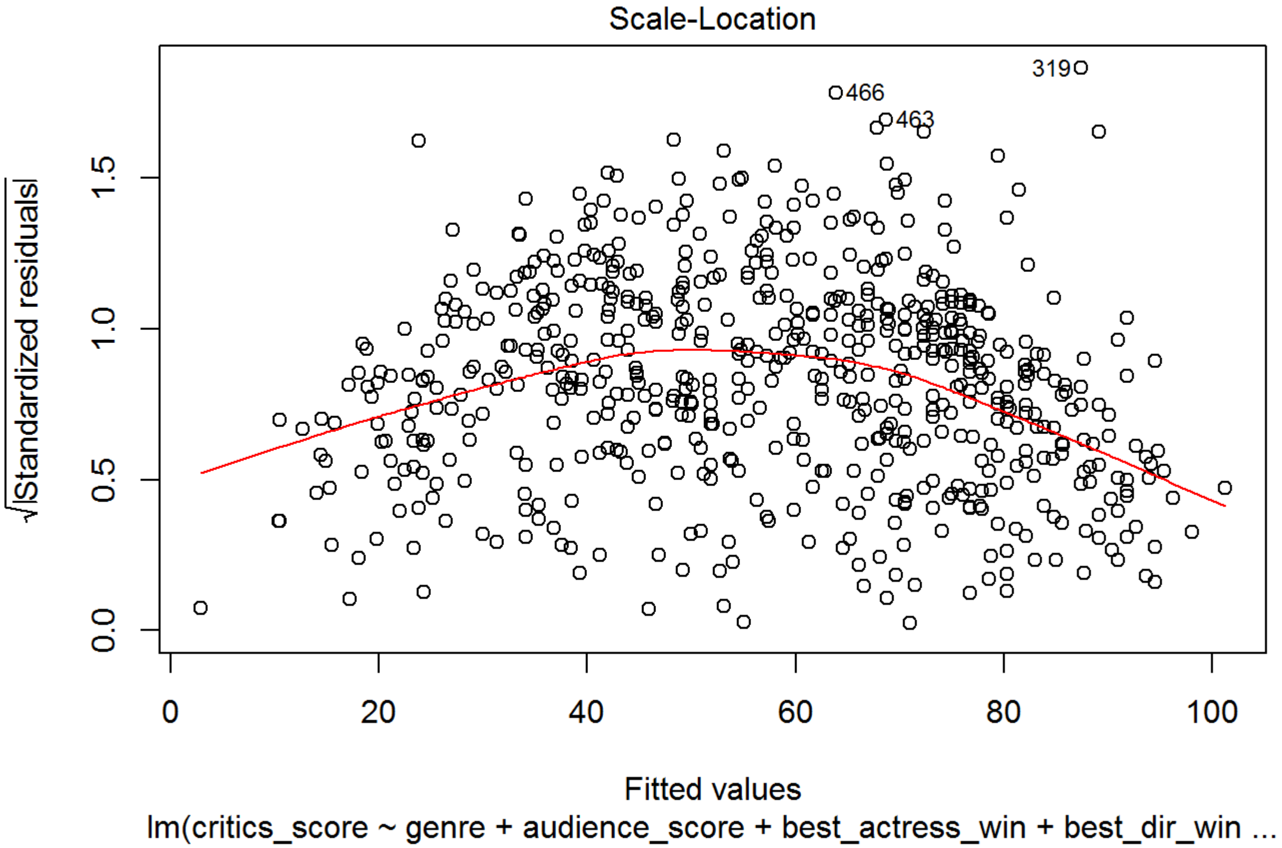
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score + best_actress_win +
##     best_dir_win, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.402 -13.340   2.395  13.974  51.662
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -6.87603    3.32876  -2.066  0.03927 *
## genreAnimation                  1.28312    6.98211   0.184  0.85425
## genreArt House & International  1.43576    5.79233   0.248  0.80431
## genreComedy                     0.45548    3.22103   0.141  0.88759
## genreDocumentary               19.62635    3.85814   5.087 4.79e-07 ***
## genreDrama                      9.89035    2.74549   3.602  0.00034 ***
## genreHorror                     9.74766    4.76151   2.047  0.04105 *
## genreMusical & Performing Arts 11.61006    6.25159   1.857  0.06375 .
## genreMystery & Suspense        10.75745    3.54751   3.032  0.00252 **
## genreOther                     11.60896    5.49933   2.111  0.03516 *
## genreScience Fiction & Fantasy 10.73336    6.96575   1.541  0.12384
## audience_score                  0.88871    0.04261  20.855  < 2e-16 ***
## best_actress_winyes             2.88377    2.50564   1.151  0.25020
## best_dir_winyes                 8.25696    3.14075   2.629  0.00877 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.58 on 637 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.525
## F-statistic: 56.25 on 13 and 637 DF,  p-value: < 2.2e-16
```

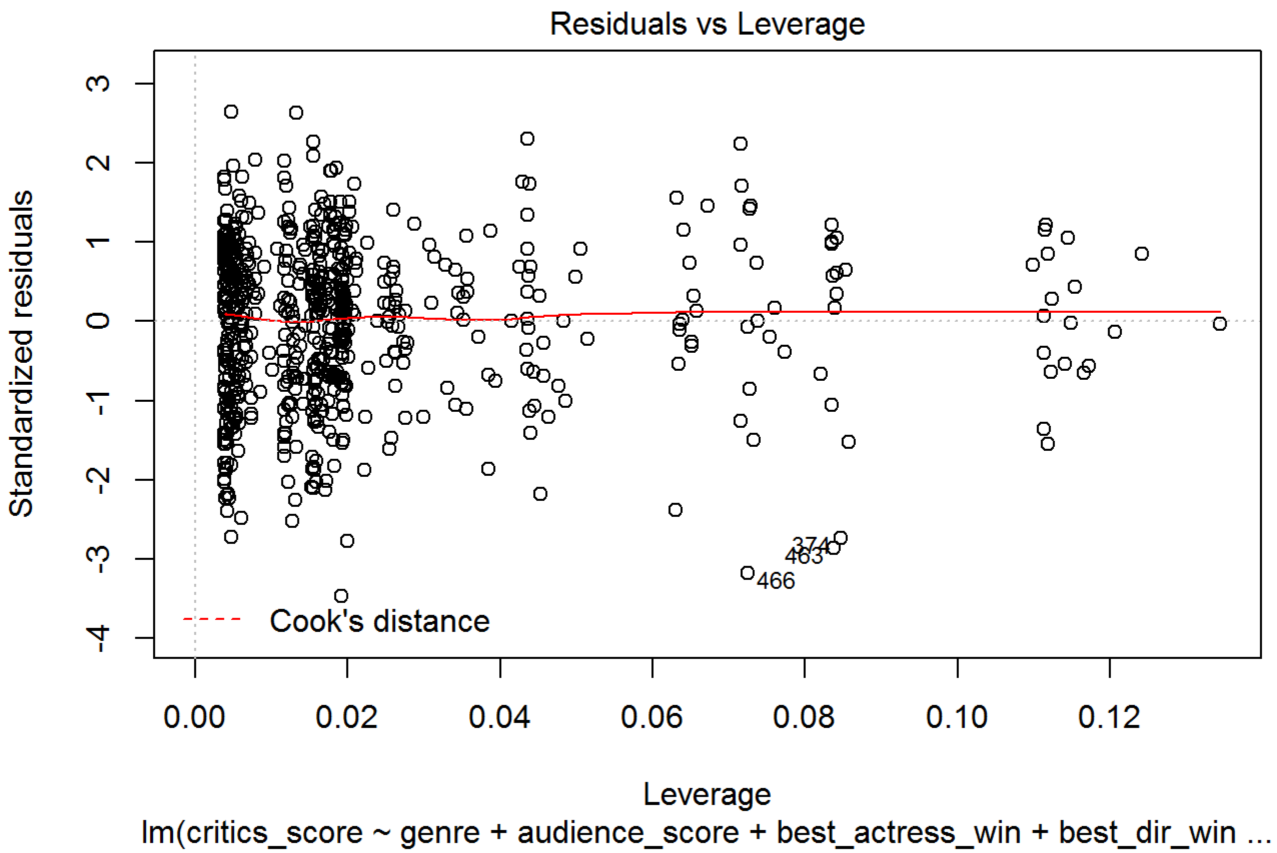
```
plot(my_model_3)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
-80
0
20
40
60
80
100
323
466
319
Fitted values
lm(critics_score ~ genre + audience_score + best_actress_win + best_dir_win ...


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
463
466
319
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score + best_actress_win + best_dir_win ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
319
466
463
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + best_actress_win + best_dir_win ...

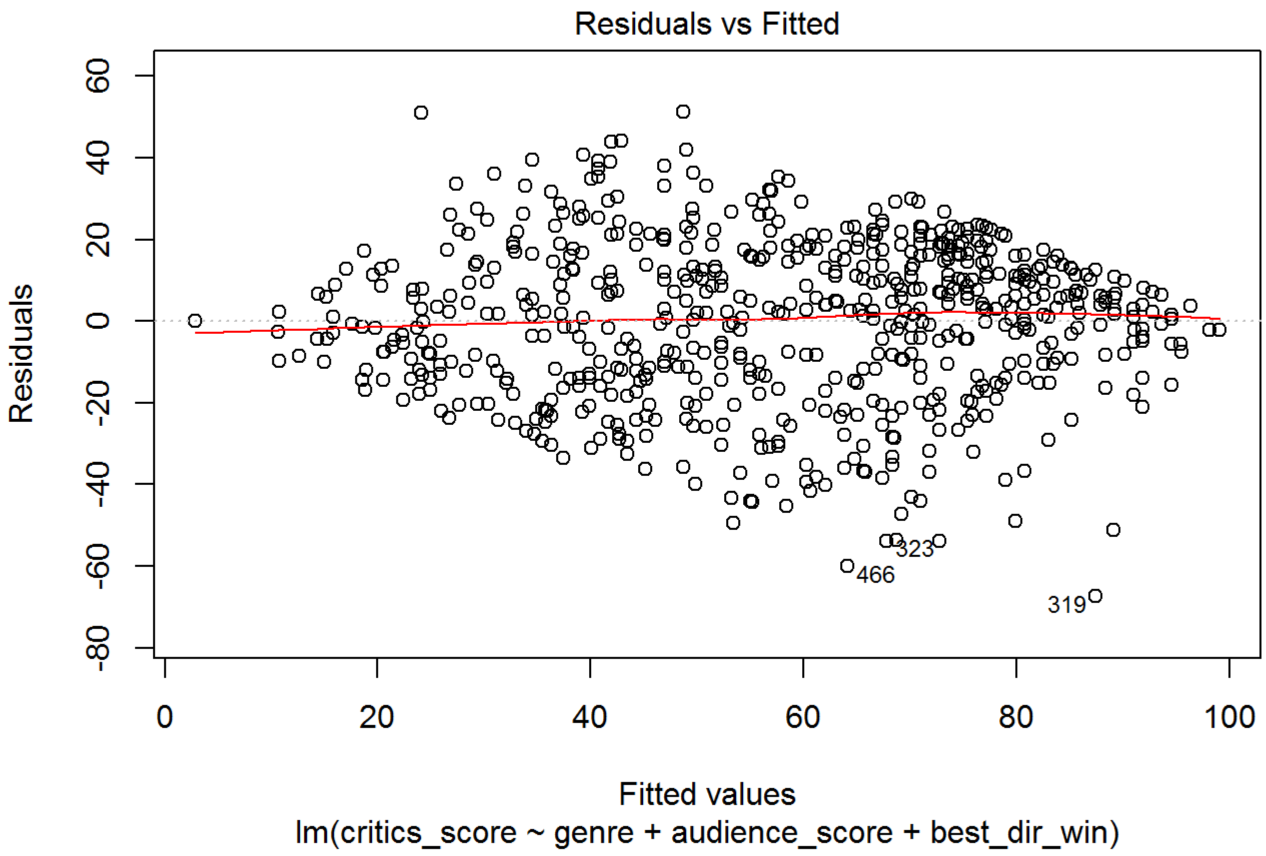


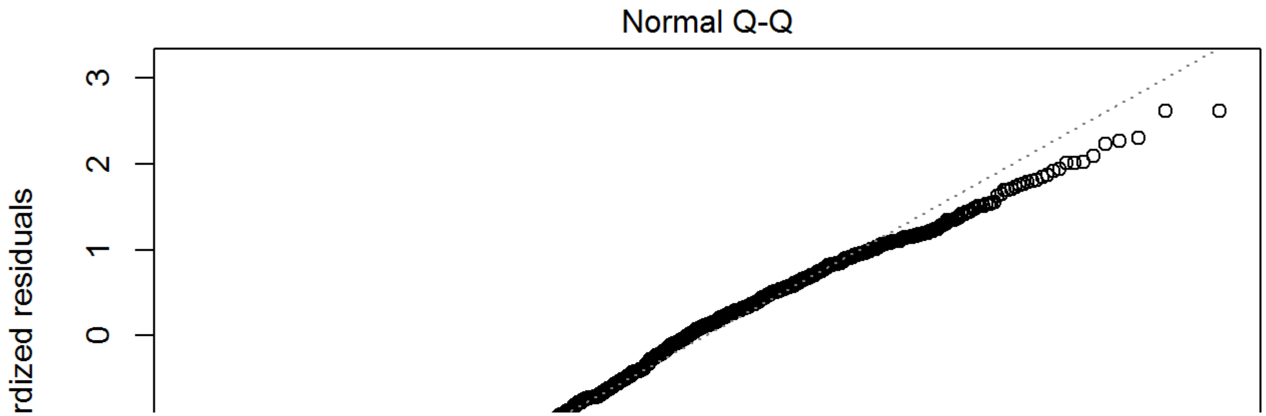
Excluding best_actress_win

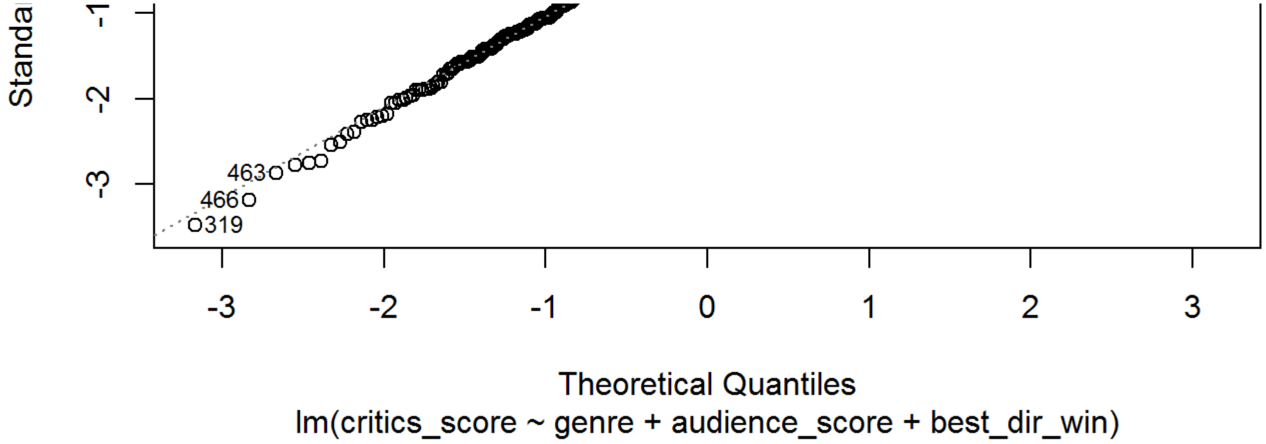
```
my_model_4 = lm( critics_score ~ genre + audience_score   + best_dir_win  , data = movies)
summary(my_model_4)
```

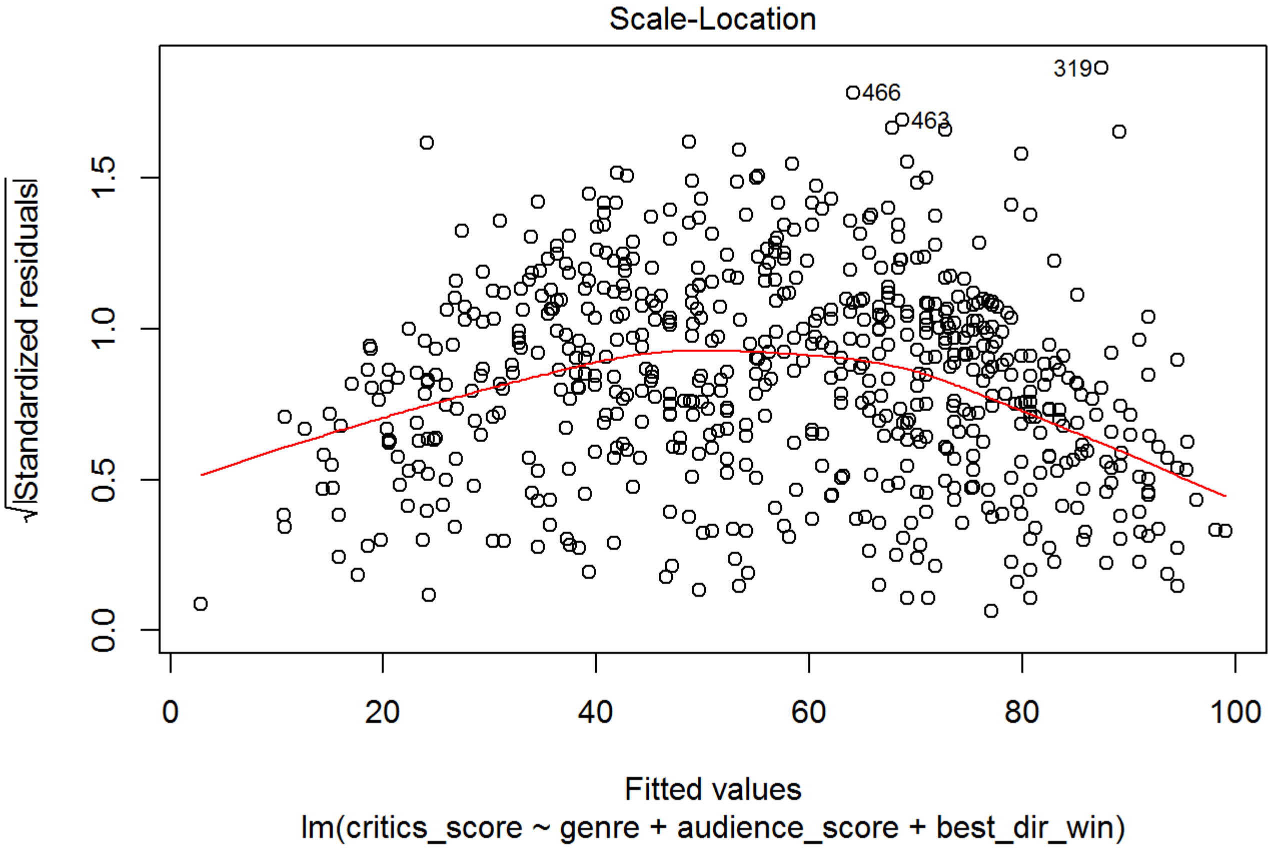
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score + best_dir_win,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.458 -13.526   2.345  13.795  51.239
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -6.93293    3.32924  -2.082 0.037700 *
## genreAnimation                  1.61214    6.97803   0.231 0.817365
## genreArt House & International  1.64914    5.79084   0.285 0.775901
## genreComedy                     0.75866    3.21106   0.236 0.813304
## genreDocumentary               19.67471    3.85889   5.099 4.52e-07 ***
## genreDrama                     10.33017    2.71945   3.799 0.000159 ***
## genreHorror                     9.75831    4.76271   2.049 0.040881 *
## genreMusical & Performing Arts 11.58424    6.25314   1.853 0.064409 .
## genreMystery & Suspense        11.23457    3.52410   3.188 0.001503 **
## genreOther                     11.95921    5.49230   2.177 0.029813 *
## genreScience Fiction & Fantasy 10.72327    6.96752   1.539 0.124291
## audience_score                  0.88948    0.04262  20.871  < 2e-16 ***
## best_dir_winyes                 8.50564    3.13411   2.714 0.006829 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.58 on 638 degrees of freedom
## Multiple R-squared:  0.5335, Adjusted R-squared:  0.5247
## F-statistic:  60.8 on 12 and 638 DF,  p-value: < 2.2e-16
```

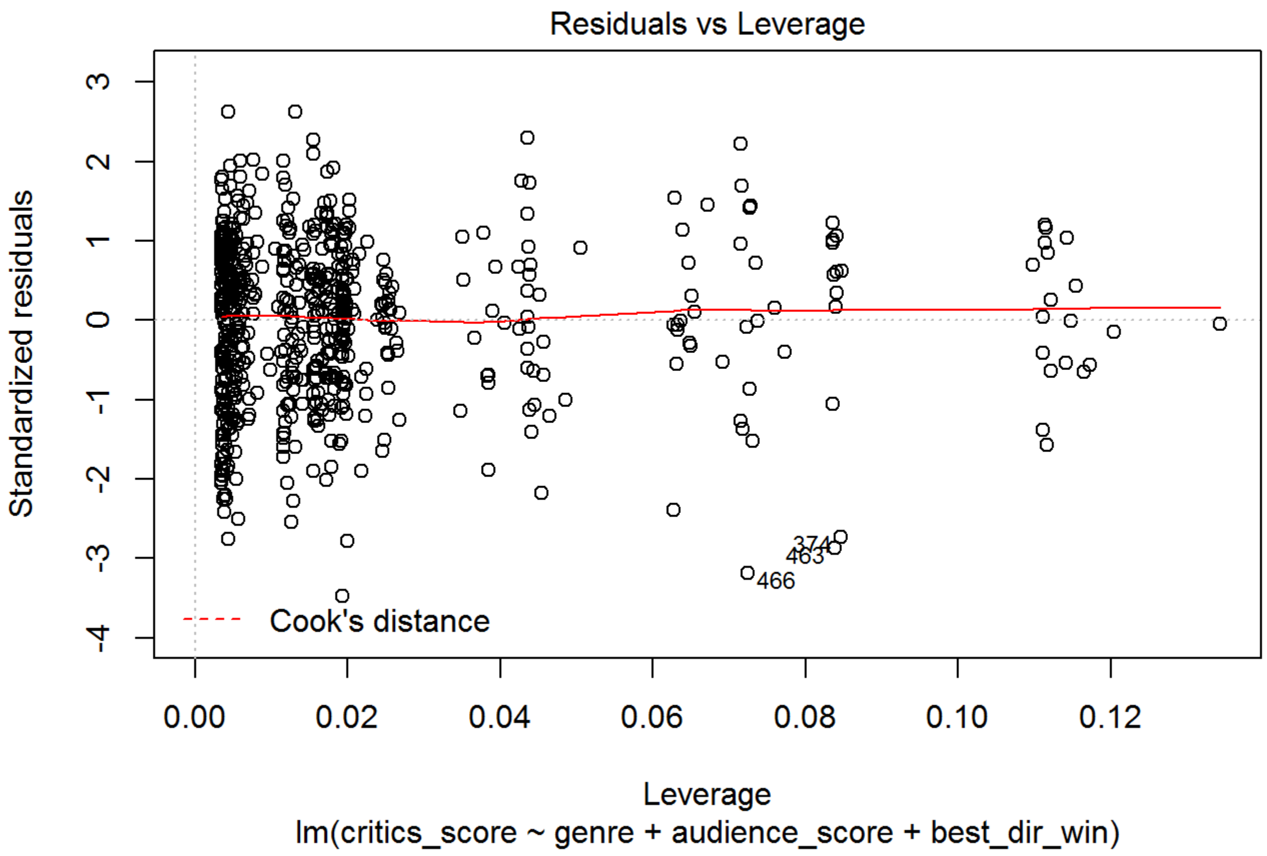
```
plot(my_model_4)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
-80
0
20
40
60
80
100
323
466
319
Fitted values
lm(critics_score ~ genre + audience_score + best_dir_win)


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
463
466
319
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score + best_dir_win)


Scale-Location
319
466
463
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score + best_dir_win)

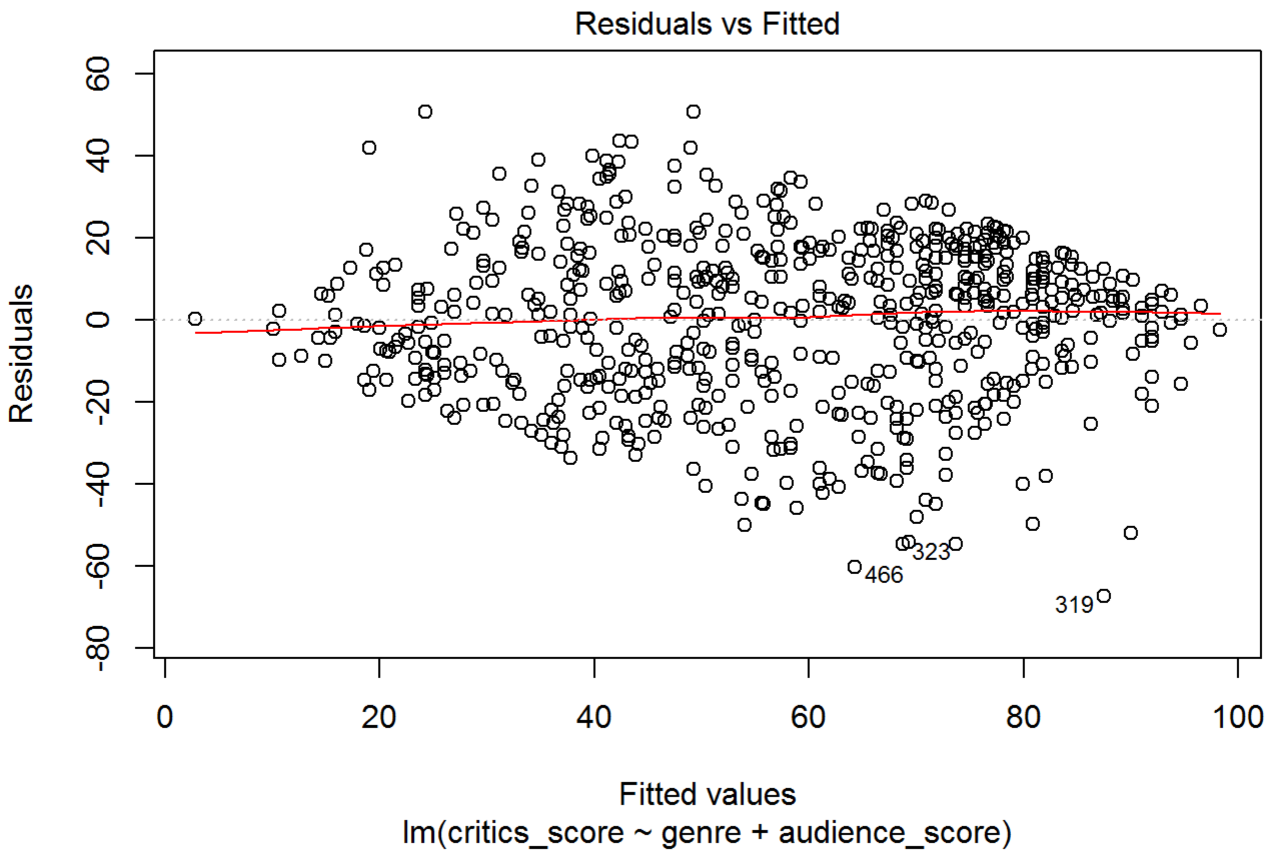


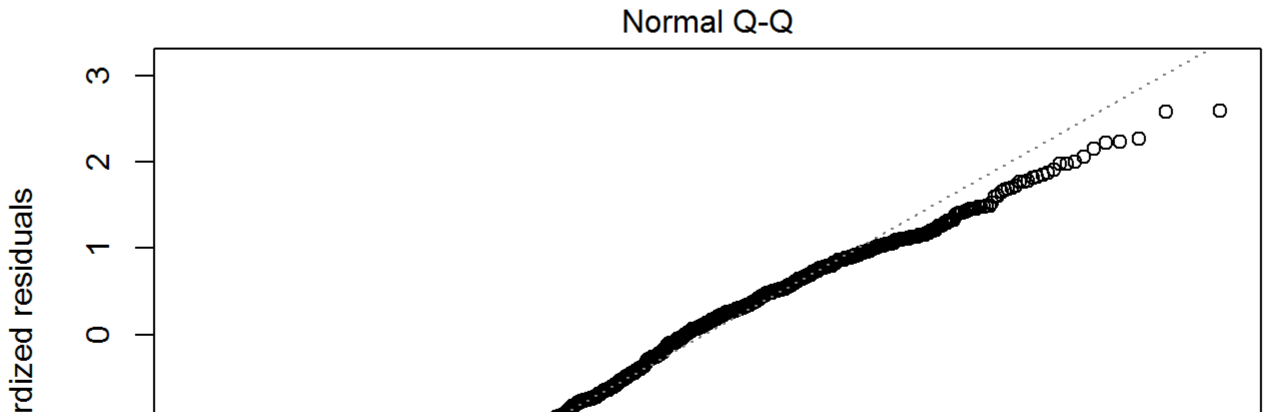
Finally , we exclude the variable best_dir_win for getting the final regression model,i.e. my_model_regre which only includes varibles critics_score, genre of movies and audience_score

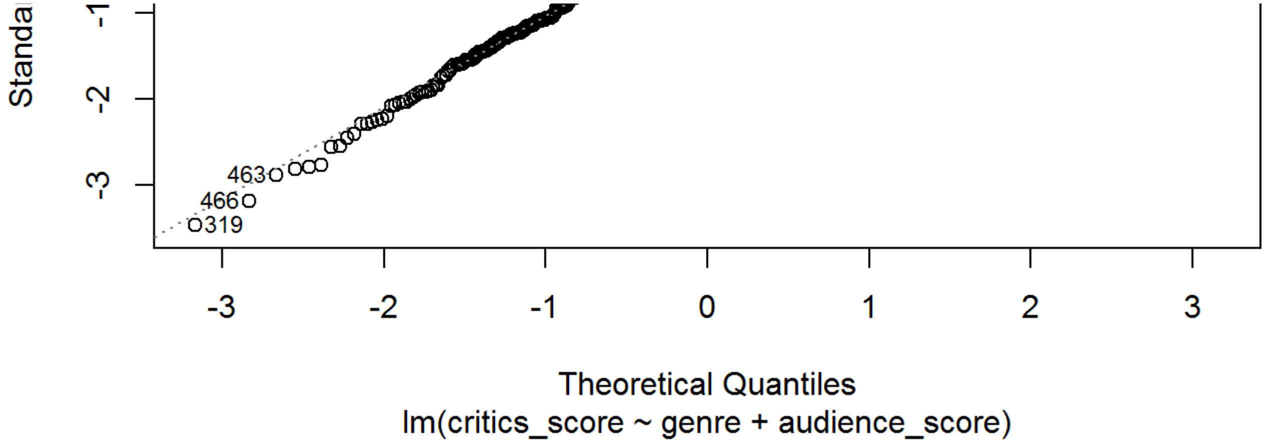
```
my_model_regre = lm( critics_score ~ genre + audience_score , data = movies)
summary(my_model_regre)
```

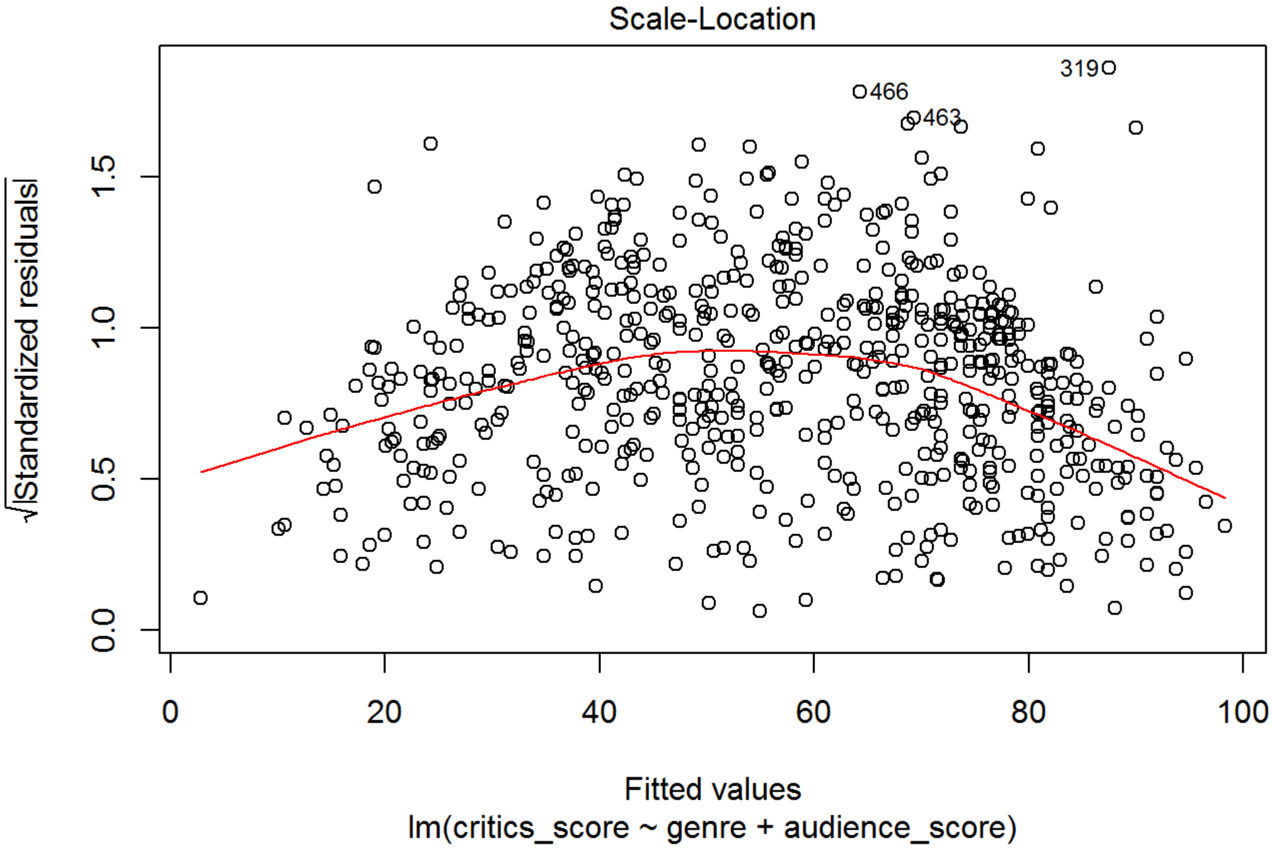
```
##
## Call:
## lm(formula = critics_score ~ genre + audience_score, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.475 -13.550   2.611  14.415  50.799
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -7.15886    3.34473  -2.140 0.032705 *
## genreAnimation                  0.96807    7.00864   0.138 0.890186
## genreArt House & International  0.98339    5.81438   0.169 0.865747
## genreComedy                     0.64412    3.22673   0.200 0.841843
## genreDocumentary               18.74773    3.86284   4.853 1.53e-06 ***
## genreDrama                     10.34282    2.73296   3.784 0.000169 ***
## genreHorror                     9.71558    4.78635   2.030 0.042786 *
## genreMusical & Performing Arts 11.40205    6.28385   1.815 0.070069 .
## genreMystery & Suspense        11.54597    3.53973   3.262 0.001166 **
## genreOther                     11.78762    5.51922   2.136 0.033080 *
## genreScience Fiction & Fantasy 11.18526    7.00004   1.598 0.110563
## audience_score                  0.90341    0.04252  21.248  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.68 on 639 degrees of freedom
## Multiple R-squared:  0.5281, Adjusted R-squared:   0.52
## F-statistic: 65.01 on 11 and 639 DF,  p-value: < 2.2e-16
```

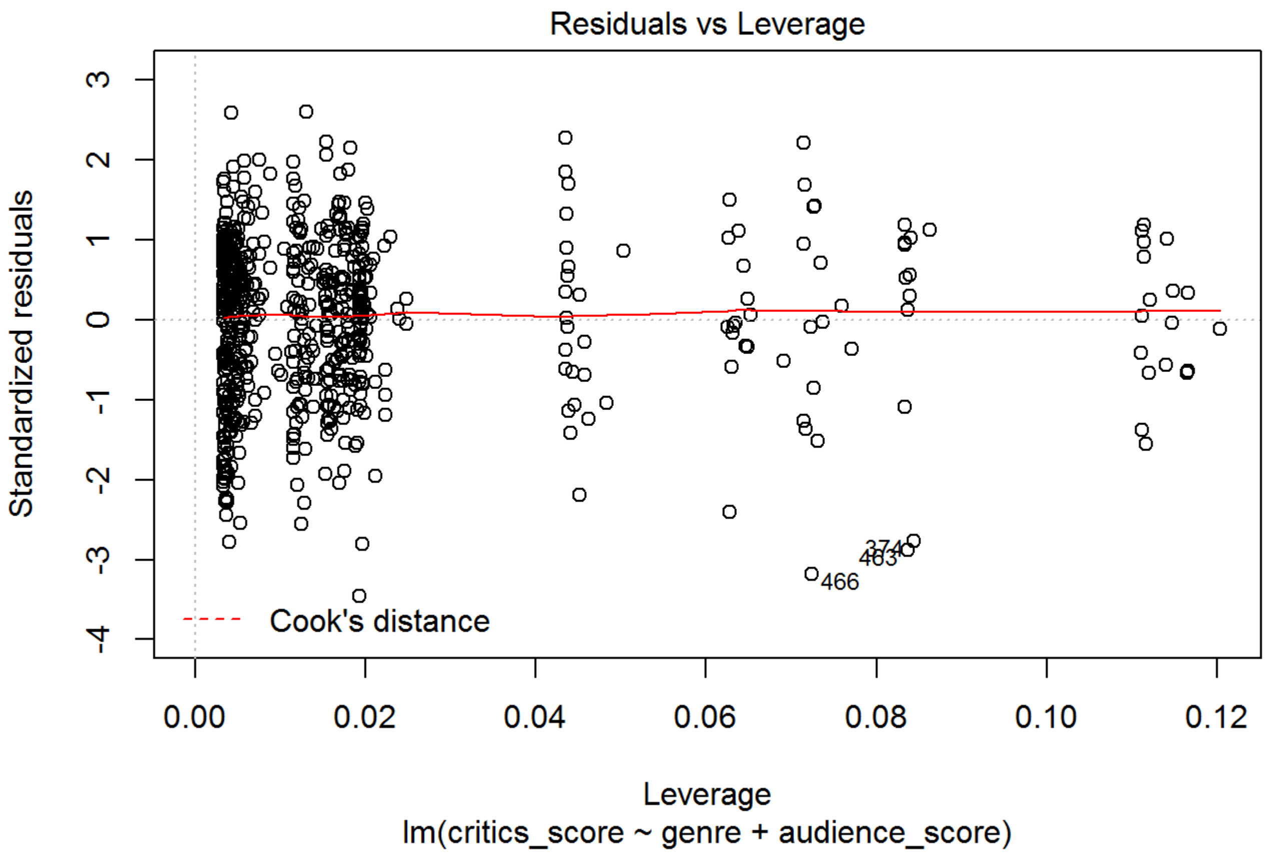
```
plot(my_model_regre)
```


Residuals vs Fitted
Residuals
60
40
20
0
-20
-40
-60
-80
323
466
319
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score)


Normal Q-Q
rdized residuals
3
2
1
0


Standa
-1
-2
-3
463
466
319
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(critics_score ~ genre + audience_score)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
319
466
463
0
20
40
60
80
100
Fitted values
lm(critics_score ~ genre + audience_score)



Part 5. Prediction

I choose Dangal movie released in 2016 to predict my model and check it out it’s reviews,i.e. score from rotten tomatoes website. I collected the information about the movie from the website “https://www.rottentomatoes.com (https://www.rottentomatoes.com/)”

1. genre = Drama, 2. score of critics = 92 and 3. score of audience = 95 .

The interval taken is confidence.

```
library(caret)
```

```
## Loading required package: lattice
```

```
predict(my_model_regre, data.frame(genre="Drama",
critics_score = 92, audience_score =95), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 89.00804 85.68761 92.32847
```

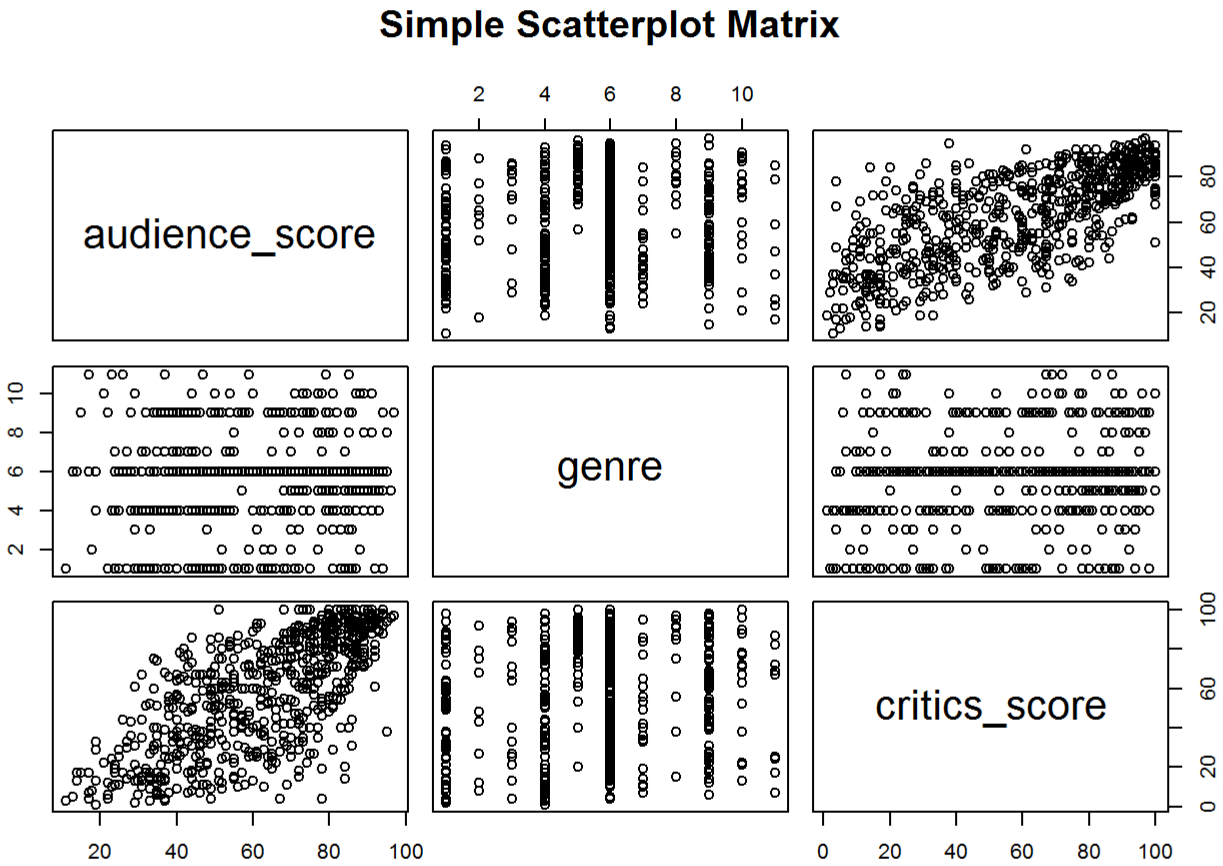
```
confint(my_model_regre)
```

```
##                                      2.5 %     97.5 %
## (Intercept)                    -13.7268581 -0.5908677
## genreAnimation                 -12.7946849 14.7308151
## genreArt House & International -10.4342144 12.4010013
## genreComedy                     -5.6921642  6.9803944
## genreDocumentary                11.1623336 26.3331210
## genreDrama                       4.9761576 15.7094916
## genreHorror                      0.3167011 19.1144538
## genreMusical & Performing Arts  -0.9374381 23.7415437
## genreMystery & Suspense          4.5950490 18.4968845
## genreOther                       0.9496189 22.6256177
## genreScience Fiction & Fantasy  -2.5606082 24.9311341
## audience_score                   0.8199193  0.9869034
```

```
par("mar")
```

```
## [1] 5.1 4.1 4.1 2.1
```

```
par(mar=c(.2,.2,.2,.2))
pairs(~audience_score +genre +critics_score, data = movies,
main = "Simple Scatterplot Matrix")
```



Result of prediction :

By this prediction modelling, we are 95% confident that the audience_score would be between 81% and 98% . After comparing this score to it's actual present score on rottentomatoes website we see that audience score is 95% on this site, and which is in the predicted range..

Part 6. Conclusion :

This project help me to understand the exploratory data analysis analysis, modeling and prediction and also made me understand the concept of the ratings of movies. According to the research question, I found that the most impact putting variables for getting the best movie review on rotten tomatoes website are the score of critics and the audience.

I found some shortcomings while in some research senerio for which I started reading more datasets and website data so I can understand and find out the researchable questions.

For the further research I recommend that we can also find that “if there is any impact on the moovie rating based on it’s release month because I realize that in some countrys outside India(my country) they have snowfall in particular months and they enjoy the outings in summers(most probably), so we can make a model which can predict this.”