Modeling and prediction for movies

Author – Varshit Dubey (CoE Pune)

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(ggthemes)
library(corrgram)
library(corrplot)
library(caTools)
library(sp)
library(raster)
library(usdm)
library(lmtest)
```

Load the data set

```
# load the data set
load("D:/Datasets/movies.Rdata")
```

Part 1: Data

The data set is comprised of 651 randomly sampled movies produced and released before 2016.

Since random sampling is used in the collection of data set and no assignment is used ,this is an observational study, not experimental.

We can only find correlation between variables and because of random sampling we can generalize the result to all the movies. We cannot find any causal relation as there is no random assignment(observational).

Part 2: Research question

By looking at the data set the basic question which arises in mind is:

What makes the movie succesfull??

Which variables contributes to the critic's rating in the movie??

Does genre, audience score affects the critic's rating of rotten tomatoes(if particular genre movie has more chance of success)??

All this question can be addressed by linear modelling..

This research question will help to search for the factors that affects the score of critics, which factors to consider while making a review..

Part 3: Exploratory data analysis

```
# Explore the data set
# Explore the first 10 observations
head(movies, 10)
```

```
## # A tibble: 10 x 32
##    title title_type genre runtime mpaa_rating studio thtr_rel_year
##    <chr> <fct>      <fct>   <dbl> <fct>       <fct>          <dbl>
##  1 Fill~ Feature F~ Drama      80 R           Indom~          2013
##  2 The ~ Feature F~ Drama     101 PG-13       Warne~          2001
##  3 Wait~ Feature F~ Come~      84 R           Sony ~          1996
##  4 The ~ Feature F~ Drama     139 PG          Colum~          1993
##  5 Male~ Feature F~ Horr~      90 R           Ancho~          2004
##  6 Old ~ Documenta~ Docu~      78 Unrated     Shcal~          2009
##  7 Lady~ Feature F~ Drama     142 PG-13       Param~          1986
##  8 Mad ~ Feature F~ Drama      93 R           MGM/U~          1996
##  9 Beau~ Documenta~ Docu~      88 Unrated     Indep~          2012
## 10 The ~ Feature F~ Drama     119 Unrated     IFC F~          2012
## # ... with 25 more variables: thtr_rel_month <dbl>, thtr_rel_day <dbl>,
## #   dvd_rel_year <dbl>, dvd_rel_month <dbl>, dvd_rel_day <dbl>,
## #   imdb_rating <dbl>, imdb_num_votes <int>, critics_rating <fct>,
## #   critics_score <dbl>, audience_rating <fct>, audience_score <dbl>,
## #   best_pic_nom <fct>, best_pic_win <fct>, best_actor_win <fct>,
## #   best_actress_win <fct>, best_dir_win <fct>, top200_box <fct>,
## #   director <chr>, actor1 <chr>, actor2 <chr>, actor3 <chr>,
## #   actor4 <chr>, actor5 <chr>, imdb_url <chr>, rt_url <chr>
```

```
tail(movies, 10)
```

```
## # A tibble: 10 x 32
##    title title_type genre runtime mpaa_rating studio thtr_rel_year
##    <chr> <fct>      <fct>   <dbl> <fct>       <fct>          <dbl>
##  1 Pina  Documenta~ Musi~     103 PG          IFC F~          2011
##  2 Capo~ Feature F~ Drama     114 R           Sony ~          2005
##  3 Dead~ Feature F~ Myst~      88 PG          Unive~          1982
##  4 Tarz~ Feature F~ Drama      88 G           Buena~          1999
##  5 Coco~ Feature F~ Drama     116 PG          Fox             1988
```

```
##  6 Deat~ Feature F~ Drama      97 PG          Geniu~          2008
##  7 Half~ Feature F~ Come~      82 R           Unive~          1998
##  8 Danc~ Feature F~ Acti~      87 R           Grind~          2008
##  9 Arou~ Feature F~ Acti~     120 PG          Buena~          2004
## 10 LOL   Feature F~ Come~      97 PG-13       Lions~          2012
## # ... with 25 more variables: thtr_rel_month <dbl>, thtr_rel_day <dbl>,
## #   dvd_rel_year <dbl>, dvd_rel_month <dbl>, dvd_rel_day <dbl>,
## #   imdb_rating <dbl>, imdb_num_votes <int>, critics_rating <fct>,
## #   critics_score <dbl>, audience_rating <fct>, audience_score <dbl>,
## #   best_pic_nom <fct>, best_pic_win <fct>, best_actor_win <fct>,
## #   best_actress_win <fct>, best_dir_win <fct>, top200_box <fct>,
## #   director <chr>, actor1 <chr>, actor2 <chr>, actor3 <chr>,
## #   actor4 <chr>, actor5 <chr>, imdb_url <chr>, rt_url <chr>
```

```
# explore the variables of movies data set
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/" ...
```

```r
# explore various statistical concepts of variables of movies data set
summary(movies)
```

```
##     title                   title_type                  genre
##  Length:651         Documentary : 55   Drama             :305
##  Class :character   Feature Film:591   Comedy            : 87
##  Mode  :character   TV Movie    :  5   Action & Adventure: 65
##                                        Mystery & Suspense: 59
##                                        Documentary       : 52
##                                        Horror            : 23
##                                        (Other)           : 60
##     runtime       mpaa_rating                                studio
##  Min.   : 39.0   G      : 19   Paramount Pictures              : 37
##  1st Qu.: 92.0   NC-17  :  2   Warner Bros. Pictures           : 30
##  Median :103.0   PG     :118   Sony Pictures Home Entertainment: 27
##  Mean   :105.8   PG-13  :133   Universal Pictures              : 23
##  3rd Qu.:115.8   R      :329   Warner Home Video               : 19
##  Max.   :267.0   Unrated: 50   (Other)                         :507
##  NA's   :1                     NA's                            :  8
##  thtr_rel_year  thtr_rel_month   thtr_rel_day    dvd_rel_year
```

```
##  Min.   :1970   Min.   : 1.00   Min.   : 1.00   Min.   :1991
##  1st Qu.:1990   1st Qu.: 4.00   1st Qu.: 7.00   1st Qu.:2001
##  Median :2000   Median : 7.00   Median :15.00   Median :2004
##  Mean   :1998   Mean   : 6.74   Mean   :14.42   Mean   :2004
##  3rd Qu.:2007   3rd Qu.:10.00   3rd Qu.:21.00   3rd Qu.:2008
##  Max.   :2014   Max.   :12.00   Max.   :31.00   Max.   :2015
##                                                 NA's   :8
##  dvd_rel_month     dvd_rel_day     imdb_rating    imdb_num_votes
##  Min.   : 1.000   Min.   : 1.00   Min.   :1.900   Min.   :   180
##  1st Qu.: 3.000   1st Qu.: 7.00   1st Qu.:5.900   1st Qu.:  4546
##  Median : 6.000   Median :15.00   Median :6.600   Median : 15116
##  Mean   : 6.333   Mean   :15.01   Mean   :6.493   Mean   : 57533
##  3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:7.300   3rd Qu.: 58301
##  Max.   :12.000   Max.   :31.00   Max.   :9.000   Max.   :893008
##  NA's   :8        NA's   :8
##          critics_rating critics_score    audience_rating audience_score
##  Certified Fresh:135    Min.   :  1.00   Spilled:275     Min.   :11.00
##  Fresh          :209    1st Qu.: 33.00   Upright:376     1st Qu.:46.00
##  Rotten         :307    Median : 61.00                   Median :65.00
##                         Mean   : 57.69                   Mean   :62.36
##                         3rd Qu.: 83.00                   3rd Qu.:80.00
##                         Max.   :100.00                   Max.   :97.00
##
##  best_pic_nom best_pic_win best_actor_win best_actress_win best_dir_win
##  no :629      no :644      no :558        no :579          no :608
##  yes: 22      yes:  7      yes: 93        yes: 72          yes: 43
##
##
##
##
##
##  top200_box   director           actor1             actor2
##  no :636    Length:651         Length:651         Length:651
##  yes: 15    Class :character   Class :character   Class :character
##             Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     actor3             actor4             actor5
##  Length:651         Length:651         Length:651
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    imdb_url           rt_url
```

```
##  Length:651         Length:651
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Now we start to dig deeper in the data set.

```
a3 <- movies %>% group_by(genre) %>% filter(!is.na(genre) ,
!is.na(imdb_rating)) %>% summarise(meanrating= mean(imdb_rating)) %>%
arrange(desc(meanrating))

a3
```

```
## # A tibble: 11 x 2
##    genre                     meanrating
##    <fct>                          <dbl>
##  1 Documentary                     7.65
##  2 Musical & Performing Arts       7.3
##  3 Drama                           6.67
##  4 Other                           6.63
##  5 Art House & International       6.61
##  6 Mystery & Suspense              6.48
##  7 Action & Adventure              5.97
##  8 Animation                       5.9
##  9 Horror                          5.76
## 10 Science Fiction & Fantasy       5.76
## 11 Comedy                          5.74
```

This result shows movies with genre "Documentary" has the highest average rating in IMDB while comedy has the lowest average rating.

```
# filtering the data according to the variables of interest.

a0 <- movies %>% group_by(genre) %>% filter(!is.na(genre),
!is.na(critics_score)) %>%
summarise(meancritic=mean(critics_score),meanaudience = mean(audience_score))
%>% mutate(diff = meanaudience- meancritic) %>% arrange(desc(diff))

a0
```

```
## # A tibble: 11 x 4
##    genre                     meancritic meanaudience    diff
##    <fct>                          <dbl>        <dbl>   <dbl>
##  1 Art House & International       51.6         64    12.4
##  2 Action & Adventure              41.4         53.8  12.4
##  3 Animation                       50.2         62.4  12.2
```

```
##  4 Comedy                          40.9         52.5  11.6
##  5 Musical & Performing Arts       76.7         80.2   3.5
##  6 Drama                           62.2         65.3   3.13
##  7 Horror                          44.0         45.8   1.87
##  8 Other                           64.9         66.7   1.81
##  9 Mystery & Suspense              54.9         55.9   1.02
## 10 Science Fiction & Fantasy       50           50.9   0.889
## 11 Documentary                     86.3         82.8  -3.60
```

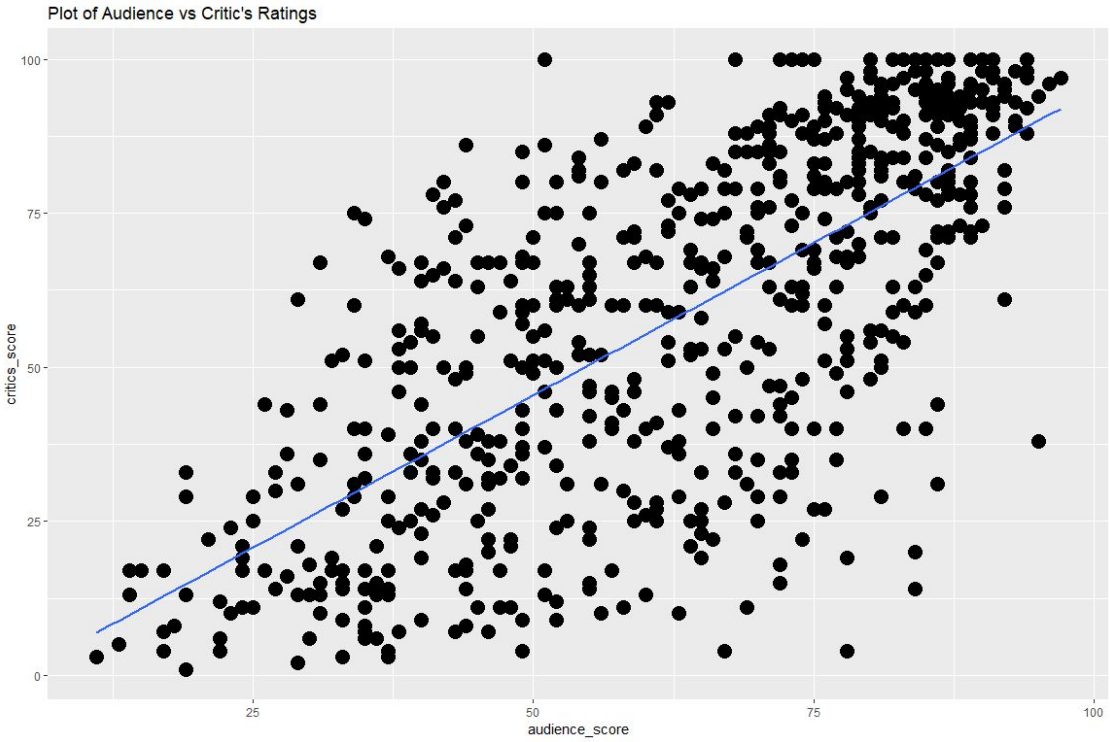
This table shows the average rating given by critics and audience based on genre. As you can see from the table there is difference in rating. So we can say that audience and critics and audience have different taste of movies.

```
# average of difference between audience and critcs rating based on genre

movies %>% mutate(diff = audience_score - critics_score) %>% group_by(genre)
%>% summarise(m = mean(diff)) %>% arrange(desc(m))
```

```
## # A tibble: 11 x 2
##    genre                          m
##    <fct>                      <dbl>
##  1 Art House & International  12.4
##  2 Action & Adventure         12.4
##  3 Animation                  12.2
##  4 Comedy                     11.6
##  5 Musical & Performing Arts   3.5
##  6 Drama                       3.13
##  7 Horror                      1.87
##  8 Other                       1.81
##  9 Mystery & Suspense          1.02
## 10 Science Fiction & Fantasy   0.889
## 11 Documentary                -3.60
```

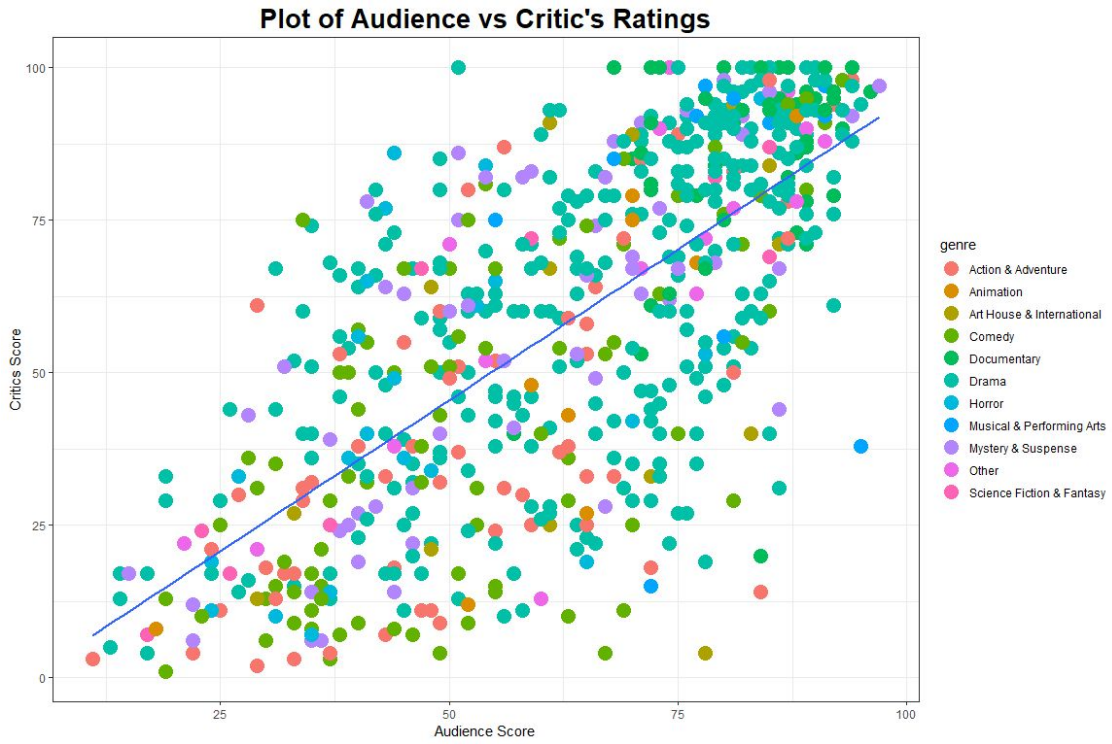
And the second table shows the differece in audience and critics rating on rotten tomatoes.The above result shows that Audience tend to like “Action and adventure” movies while Critics give them low ratings on average.

```
# Looking at the data by some visualization..

ggplot(data= movies , aes(x= audience_score ,y= critics_score)) +
  geom_point(size = 5) + geom_smooth(method= "lm",se= FALSE) + ggtitle("Plot
of Audience vs Critic's Ratings")
```

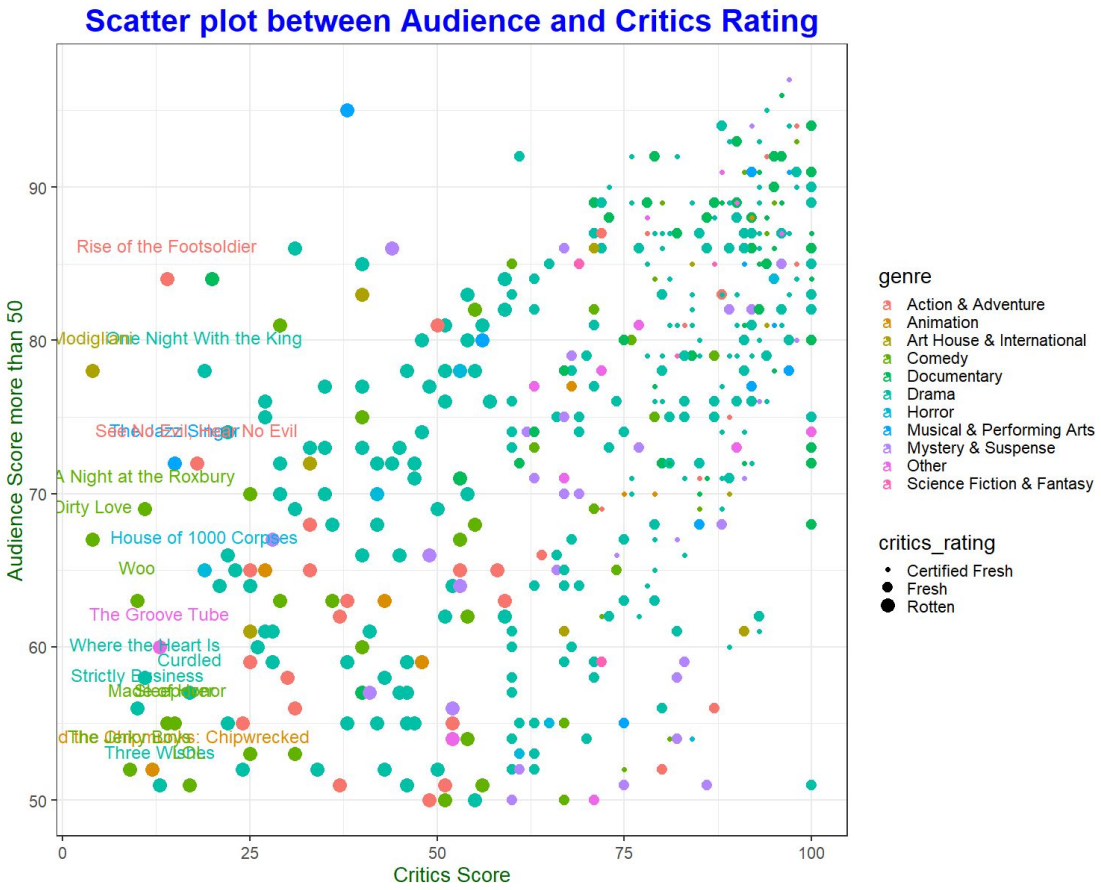


The above plot shows there is positive linear relation between audience and critic scores. Let's make this plot more attractive.

```
ggplot(data= movies , aes(x= audience_score ,y= critics_score)) +
geom_point(mapping = aes(colour = genre), size = 5) + geom_smooth(method=
"lm",se= FALSE) + ggtitle("Plot of Audience vs Critic's Ratings") +
  theme_bw() + xlab("Audience Score") +ylab("Critics Score") +
  theme(plot.title = element_text(size = 20,
                                  face = "bold",
                                  hjust = 0.5))
```




```
mod_movies <- movies %>% filter(audience_score >= 50)

ggplot(mod_movies, aes(critics_score, audience_score, color = genre)) +
geom_point(aes(size = critics_rating), fill = 0.7) + geom_text(aes(label =
ifelse(audience_score >= 50 & critics_score < 20, as.character(title),'')),

hjust = 0.5,vjust = -2), size = 6, face = "bold") + xlab("Critics Score") +
  ylab("Audience Score more than 50")  +
  ggtitle("Scatter plot between Audience and Critics Rating") +
  theme_bw(base_size = 20) +
  theme(plot.title = element_text(size = 30, face = "bold", hjust = 0.5,
colour = "Blue"),
        axis.title.x = element_text(size = 20, colour = "Dark Green"),
        axis.title.y = element_text(size = 20, colour = "Dark Green"))

## Warning: Ignoring unknown parameters: face

## Warning: Using size for a discrete variable is not advised.
```



I selected rows whose audience_score is more than 50 and stored the new data frame.

And I plot the scatterplot, between critics_score an audience_score, and also added size and colour to make it more beautiful. Ignore the warning.

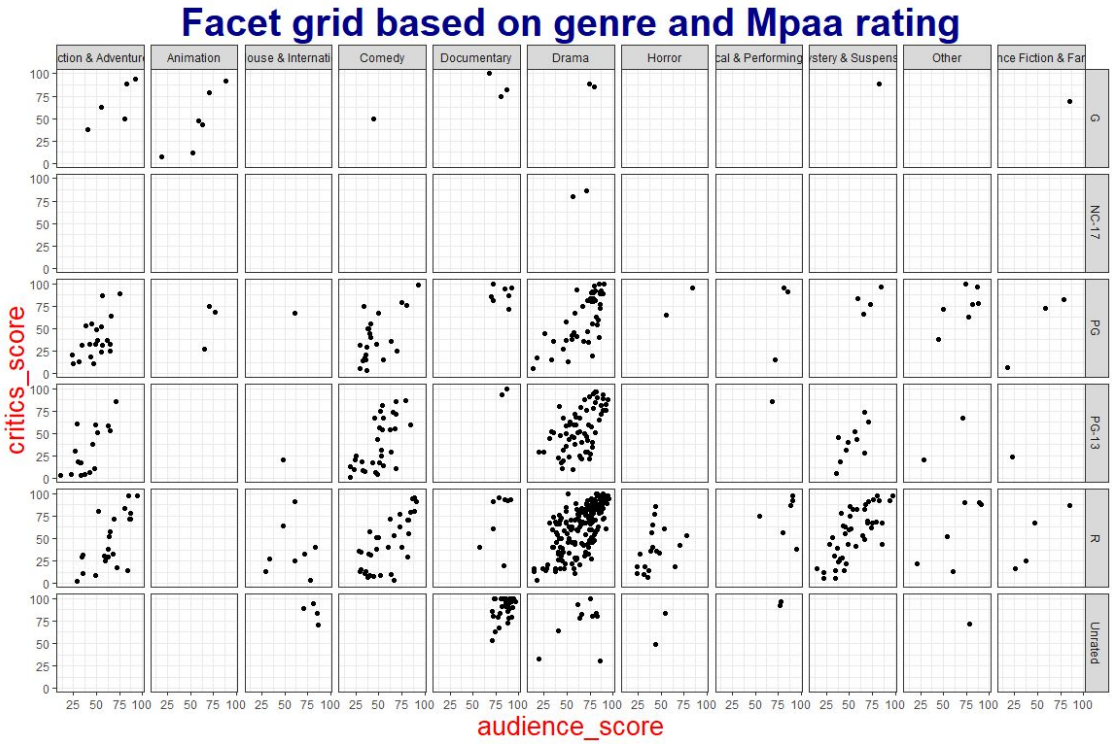
Conclusion - Surely there are some movies whose critics_score is less than 20 but audience rated it more than 50. I highlighted these movies. Interesting.

Audience rated some movies even more than 80 while critics rated it less than even 25, for ex. see the movie “Rise of the Footsoldier”.

Now we try to make a scatterplot which shows the plot based on individual levels of categorical variables, like a grid.

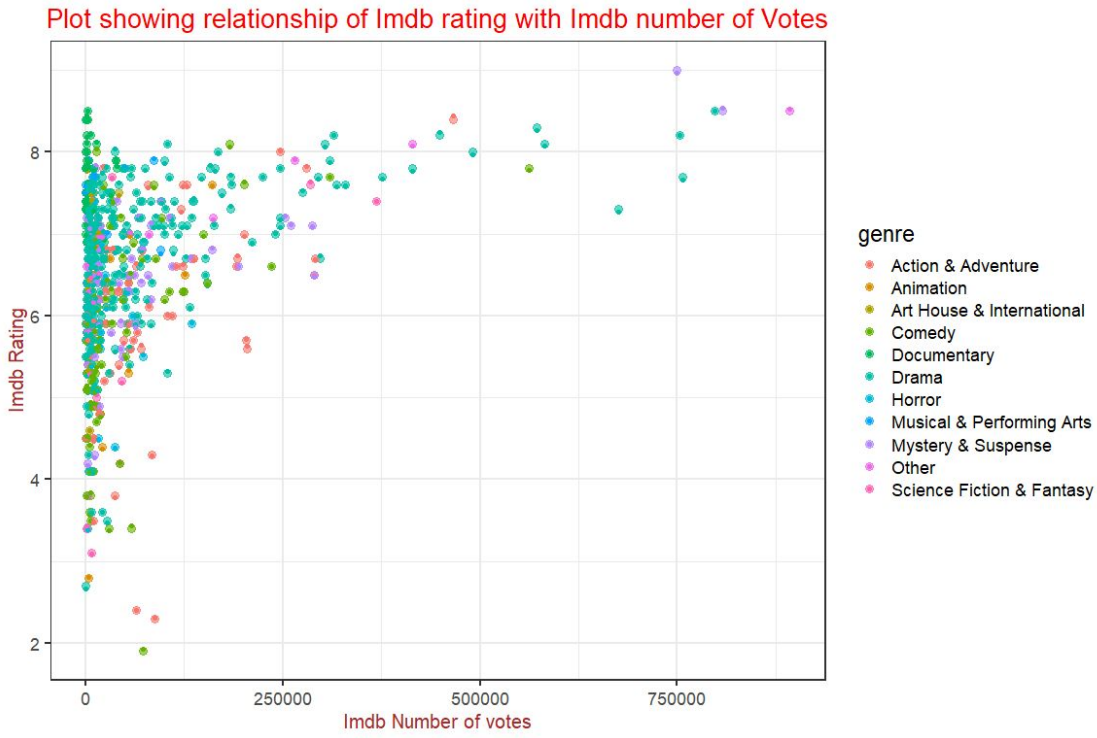
```
ggplot(movies) + geom_point(mapping = aes(audience_score, critics_score)) +
  facet_grid(mpaa_rating~genre) +
  labs(title = "Facet grid based on genre and Mpaa rating") +
  theme_bw() +
  theme(axis.title.x = element_text(size = 20, colour = "Red"),
        axis.title.y = element_text(size = 20, colour = "Red"),
        plot.title = element_text(size = 30, hjust = 0.5, colour = "Dark
```

```
Blue", face = "bold"),
        axis.title.x.top = element_text(size = 10))
```



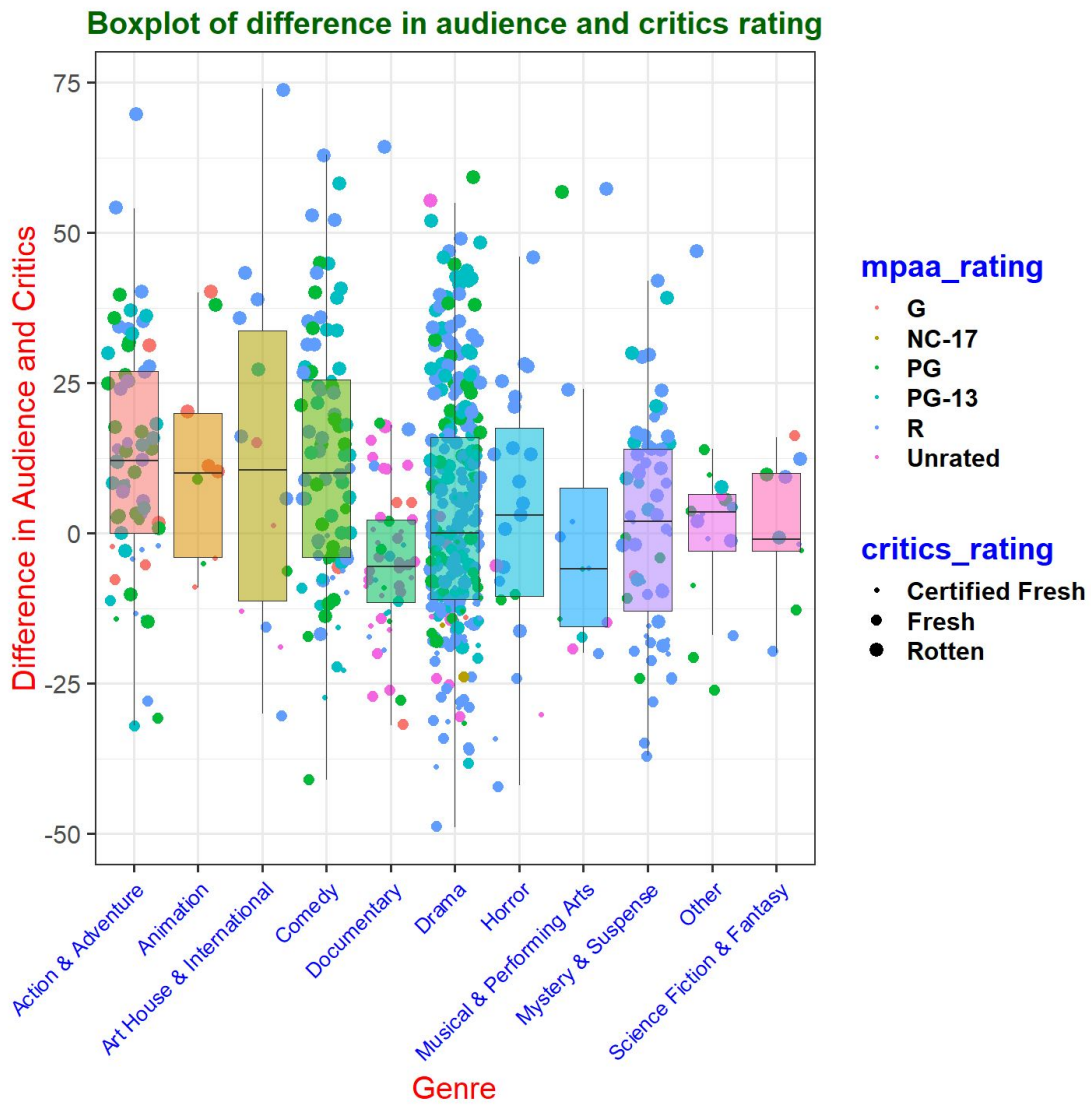
Now let's try to make scatter plot with another variables..

```
ggplot(data= movies, aes(x= imdb_num_votes, y = imdb_rating)) +
  geom_point(aes(colour = genre), size = 3, alpha = 0.6) +
  geom_jitter(aes(colour = genre)) + xlab("Imdb Number of votes") +
  ggtitle("Plot showing relationship of Imdb rating with Imdb number of
Votes")+
  ylab("Imdb Rating") +
  theme_bw(base_size = 17)  +
  theme(plot.title = element_text(hjust = 0.5, size = 20, colour = "Red"),
        axis.title.x = element_text(size = 15, colour = "Brown"),
        axis.title.y = element_text(size = 15, colour = "Brown"))
```



The plot is very dense for imdb_num_vote less than 125,000. Greater than 125,000 the points are very scattered. Also one can infer that if the Imdb number of votes is greater than 300,000 or 500,000, chances are that the Imdb rating of that movie is greater than 7.5 or 8, so greater is the no. of votes more is the imdb_rating.

```
movies %>%
  mutate(diff = audience_score - critics_score) %>%
  ggplot(aes(genre,diff))  +
  geom_jitter(aes(colour = mpaa_rating,size = critics_rating)) +
  geom_boxplot(aes(fill = genre, alpha = 0.1),show.legend = F,outlier.shape =
NA) +
  theme_bw(base_size = 30) + labs(y = "Difference in Audience and Critics", x
= "Genre", title = "Boxplot of difference in audience and critics rating") +
  theme(plot.title = element_text(size = 30, face = "bold", hjust = 0.5,
colour = "Dark Green"),
        axis.title.x = element_text(size = 30, colour = "Red"),
        axis.title.y = element_text(size = 30, colour = "Red"),
        axis.text.x = element_text(angle = 45, colour = "Blue", hjust = 1,
size = 20),
        legend.key.size = unit(2,"line"),
        legend.title = element_text(colour = "blue", face = "bold"),
        legend.text = element_text(face = "bold"))

## Warning: Using size for a discrete variable is not advised.
```



The above boxplot shows the distribution of observations of difference in audience and critics score, based on different genres. We can see that audience tend to score the movie more than critics for most of the genre as their the median of many boxplots is more than 0, but for some genres like "Documentary" the median is less than 0, means critics tend to score more than audience in this case.

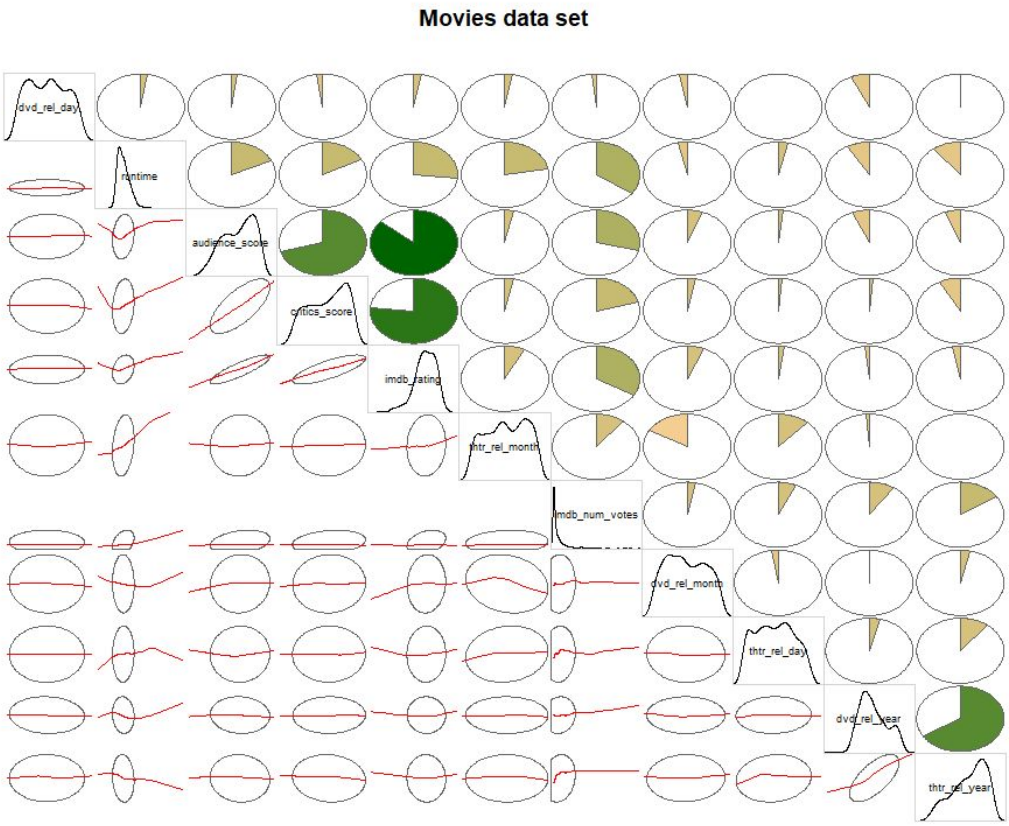
```
# Now we find correlation coefficient between all the numeric variables of
the data set.

# Removing missing values

dat <- movies %>%
  filter(!is.na(runtime), !is.na(dvd_rel_day), !is.na(dvd_rel_month),
!is.na(dvd_rel_year))
```

```r
# Grab only numeric columns
num.cols <- sapply(dat, is.numeric)

# Filter to numeric columns for correlation
cor.data <- cor(dat[,num.cols])

cor.data
```

```
##                    runtime thtr_rel_year thtr_rel_month thtr_rel_day
## runtime         1.00000000 -0.1204212193   0.2260200843  0.041135107
## thtr_rel_year  -0.12042122  1.0000000000  -0.0001711866  0.117247359
## thtr_rel_month  0.22602008 -0.0001711866   1.0000000000  0.119844827
## thtr_rel_day    0.04113511  0.1172473588   0.1198448273  1.000000000
## dvd_rel_year   -0.08190171  0.6599933006  -0.0114110547  0.043742732
## dvd_rel_month  -0.03330926  0.0390151651  -0.1667916115 -0.029343784
## dvd_rel_day     0.02423522 -0.0045649379   0.0274612137  0.003124357
## imdb_rating     0.26688085 -0.0415960198   0.0805781895  0.027618204
## imdb_num_votes  0.34668581  0.1518840288   0.1075681877  0.068603984
## critics_score   0.16777257 -0.0935587986   0.0387572967  0.017671359
## audience_score  0.17901199 -0.0611766417   0.0399363579  0.022236545
##                dvd_rel_year dvd_rel_month  dvd_rel_day imdb_rating
## runtime        -0.081901713  -0.033309263  0.024235224  0.26688085
## thtr_rel_year   0.659993301   0.039015165 -0.004564938 -0.04159602
## thtr_rel_month -0.011411055  -0.166791611  0.027461214  0.08057819
## thtr_rel_day    0.043742732  -0.029343784  0.003124357  0.02761820
## dvd_rel_year    1.000000000  -0.004092308 -0.069067849 -0.01671502
## dvd_rel_month  -0.004092308   1.000000000 -0.028817615  0.06727135
## dvd_rel_day    -0.069067849  -0.028817615  1.000000000  0.02611942
## imdb_rating    -0.016715018   0.067271350  0.026119422  1.00000000
## imdb_num_votes  0.094585300   0.029719263 -0.015977807  0.33440450
## critics_score   0.014030091   0.033072116 -0.024931612  0.76156593
## audience_score -0.063757813   0.058641662  0.021236705  0.86271975
##                imdb_num_votes critics_score audience_score
## runtime             0.34668581    0.16777257     0.17901199
## thtr_rel_year       0.15188403   -0.09355880    -0.06117664
## thtr_rel_month      0.10756819    0.03875730     0.03993636
## thtr_rel_day        0.06860398    0.01767136     0.02223654
## dvd_rel_year        0.09458530    0.01403009    -0.06375781
## dvd_rel_month       0.02971926    0.03307212     0.05864166
## dvd_rel_day        -0.01597781   -0.02493161     0.02123670
## imdb_rating         0.33440450    0.76156593     0.86271975
## imdb_num_votes      1.00000000    0.20887599     0.29178550
## critics_score       0.20887599    1.00000000     0.70024602
## audience_score      0.29178550    0.70024602     1.00000000
```

```r
# Making correlation plot
```

```
corrgram(movies, order=TRUE, main="Movies data set",
         lower.panel=corrgram::panel.ellipse,
         upper.panel=panel.pie, diag.panel=panel.density,
         col.regions=colorRampPalette(c("darkgoldenrod4", "burlywood1",
                                        "darkkhaki", "darkgreen")))
```

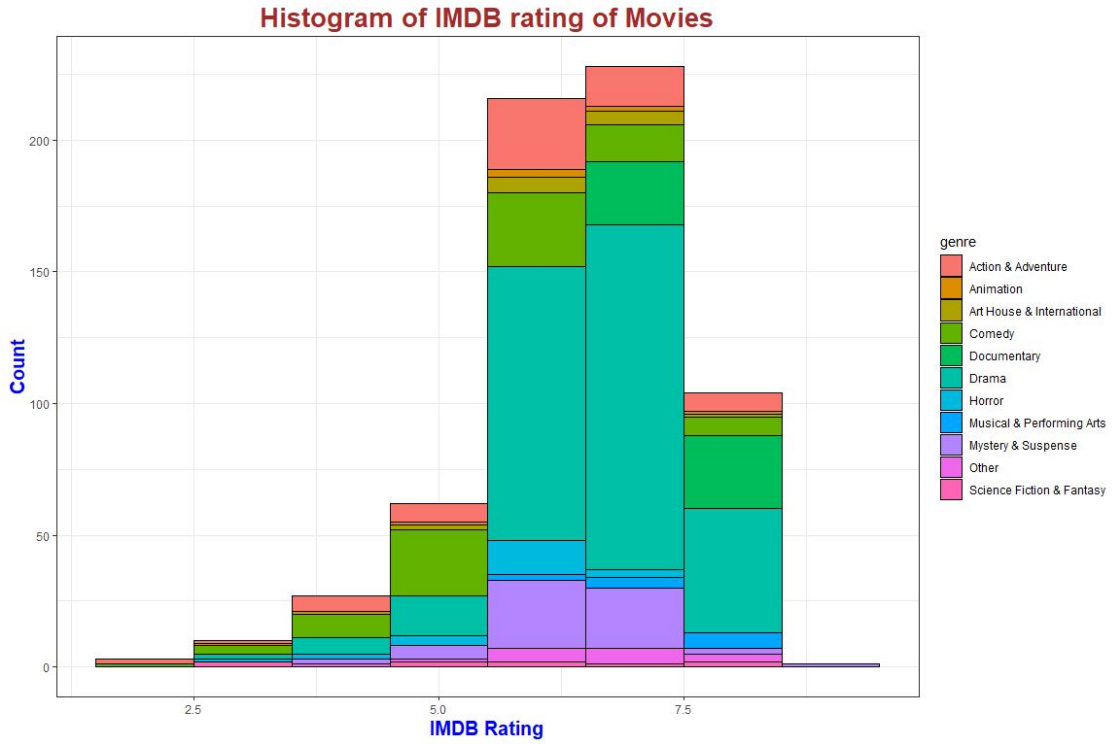


The plot shown in the figure visualises the big correlation table we made, now it becomes easier to make conclusions regarding which variables are more correlated to each other and which are not as it is difficult to make conclusions based on just observing the correlation matrix. The more a box is blue, more correlated are the 2 variables which made that box. As expected, audience score, critics_score, imdb_rating, are somewhat more correlated to each other than the rest.

Now we make a histogram of imdb rating of movies and see the type of distribution.

```
ggplot(data = movies, aes(x = imdb_rating)) +
  geom_histogram(mapping = aes(fill = genre), colour = "black", binwidth = 1)
+
  theme_bw() +
  labs(title = "Histogram of IMDB rating of Movies", x = "IMDB Rating", y =
"Count") +
  theme(plot.title = element_text(size = 20, face = "bold", color = "Brown",
```

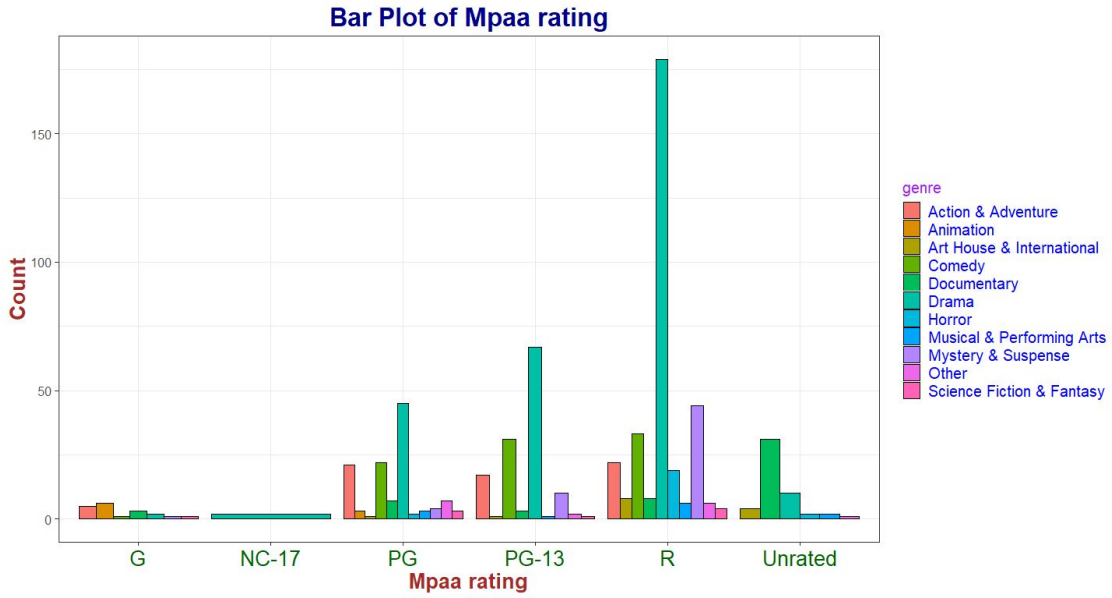
```
hjust = 0.5),
        axis.title.x = element_text(face = "bold", color = "Blue", size =
15),
        axis.title.y = element_text(face = "bold", color = "blue", size =
15))
```



In this histogram we also segregated the count by different genre, like in a range of imdb_rating, which genre has how many movie. We can see that the distribution of imdb_rating is slightly left skewed.

Now let's see another visual..

```
ggplot(movies, aes(mpaa_rating)) +
  geom_bar(aes(fill = genre), color = "black", position = "dodge") +
  labs(title = "Bar Plot of Mpaa rating", y = "Count", x = "Mpaa rating") +
  theme_bw(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, size = 25, face = "bold",
color = "Dark Blue"),
        axis.title.x = element_text(size = 20, face = "bold", colour =
"Brown"),
        axis.title.y = element_text(size = 20, face = "bold", colour =
"Brown"),
        axis.text.x = element_text(size = 20, colour = "Dark Green"),
        legend.title = element_text(colour = "Purple", size = 15),
        legend.text = element_text(size = 15, colour = "Blue"))
```



This bar plot shows count of various mpaa rating of movies with genre, one can see from the graph that "Horrer" movies are mostly rated "R", most of the movies in our data set are "R" rated, with "G" category the least.

Part 4: Modeling

To make a model, I am going to predict the critcs score of rotten tomatoes.

I am not using the variables "actor1" to "actor5" as they are statistically insignificant variables, also I am not adding release date variables as they are also statistically insignificant variables. Though I think year and month can have significant role to play in the model as this data set contains information of movies as old as 1970's and I think there is a steady transformation in the critic score and their thinking for a movie from 1970 to 2016.

I am not including "title" as the movie title name is of no use, also I removed "studio", though it can have some effect on our model, but there are 211 studio in our data set, some studio has 1 or 2 movies some have even more than 30, it will only add confusion to our model, so I removed studio variable. You can add studio in your model if you want.

```
# data preprocessing step
# Removing some variables which are statistically insignificant.
class(movies)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
mod_data <- movies[,-c(25:32)]
mod_data <- mod_data[,-c(1,6,9,12)]
```

```
# Removing or replacing Missing values
mod_data$runtime <- ifelse(is.na(mod_data$runtime), mean(mod_data$runtime,
na.rm = T), mod_data$runtime)

mod_data <- mod_data %>% filter(!is.na(dvd_rel_year), !is.na(dvd_rel_month))

# Dividing the data into training and test set

# set.seed(123)
# split = sample.split(mod_data$critics_score, SplitRatio = 0.70)
# training_set = subset(mod_data,split == T)
# test_set = subset(mod_data, split == F)

# checking vif for all numeric values..

mod_data1 <- select_if(mod_data, is.numeric)

vif(data.frame(mod_data1[,-8]))
```

```
##        Variables      VIF
## 1        runtime 1.273205
## 2  thtr_rel_year 1.850518
## 3 thtr_rel_month 1.089128
## 4   dvd_rel_year 1.792483
## 5  dvd_rel_month 1.040576
## 6    imdb_rating 4.217520
## 7 imdb_num_votes 1.289179
## 8 audience_score 4.018047
```

Now I will make the model for all the available variables in the modified data set..

```
m1 <- lm(critics_score~., data = mod_data)
summary(m1)
```

```
##
## Call:
## lm(formula = critics_score ~ ., data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.498  -7.261  -0.111   7.318  29.747
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.288e+01  2.112e+02   0.203  0.83921
## title_typeFeature Film        -9.605e+00  4.103e+00  -2.341  0.01955 *
## title_typeTV Movie            -5.303e+00  6.457e+00  -0.821  0.41185
## genreAnimation                -6.408e-01  4.397e+00  -0.146  0.88417
```

```
## genreArt House & International -6.443e+00  3.499e+00  -1.842  0.06600 .
## genreComedy                     1.093e+00  1.872e+00   0.584  0.55967
## genreDocumentary               -5.721e+00  4.416e+00  -1.296  0.19558
## genreDrama                      2.335e+00  1.657e+00   1.409  0.15947
## genreHorror                     2.598e+00  2.781e+00   0.934  0.35064
## genreMusical & Performing Arts  1.071e+00  3.803e+00   0.282  0.77834
## genreMystery & Suspense         1.203e+00  2.106e+00   0.571  0.56804
## genreOther                      1.166e+00  3.165e+00   0.368  0.71279
## genreScience Fiction & Fantasy -9.493e-01  4.146e+00  -0.229  0.81896
## runtime                         1.547e-02  2.821e-02   0.548  0.58364
## mpaa_ratingNC-17               -1.682e+00  8.405e+00  -0.200  0.84143
## mpaa_ratingPG                  -2.836e+00  3.134e+00  -0.905  0.36589
## mpaa_ratingPG-13               -5.110e+00  3.280e+00  -1.558  0.11972
## mpaa_ratingR                   -4.560e+00  3.154e+00  -1.446  0.14872
## mpaa_ratingUnrated             -6.698e-01  3.680e+00  -0.182  0.85564
## thtr_rel_year                  -1.461e-01  6.116e-02  -2.389  0.01722 *
## thtr_rel_month                 -6.259e-02  1.300e-01  -0.482  0.63031
## dvd_rel_year                    1.404e-01  1.315e-01   1.067  0.28618
## dvd_rel_month                  -2.383e-02  1.331e-01  -0.179  0.85793
## imdb_rating                     8.923e+00  9.276e-01   9.620  < 2e-16 ***
## imdb_num_votes                 -1.431e-05  5.287e-06  -2.707  0.00699 **
## critics_ratingFresh            -7.967e+00  1.390e+00  -5.733 1.55e-08 ***
## critics_ratingRotten           -4.121e+01  1.519e+00 -27.124  < 2e-16 ***
## audience_ratingUpright         -7.592e-01  1.807e+00  -0.420  0.67457
## audience_score                  1.770e-02  6.449e-02   0.274  0.78389
## best_pic_nomyes                 1.402e+00  2.926e+00   0.479  0.63213
## best_pic_winyes                 1.382e-02  5.119e+00   0.003  0.99785
## best_actor_winyes               1.050e-01  1.322e+00   0.079  0.93669
## best_actress_winyes             9.754e-01  1.455e+00   0.670  0.50295
## best_dir_winyes                 9.826e-01  1.906e+00   0.516  0.60631
## top200_boxyes                   1.517e+00  3.104e+00   0.489  0.62527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.95 on 608 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8508
## F-statistic: 108.7 on 34 and 608 DF,  p-value: < 2.2e-16
```

Well there are many insignificant variables in the data set.. There are many approaches to get to the final solution, like using step, chacking p value, Adjusted R- Squared method, checking vif. First, I'm gonna use step function to find the final set of variables in our final model..

```
m2 <- step(m1)
```

```
## Start:  AIC=3111.51
## critics_score ~ title_type + genre + runtime + mpaa_rating +
##     thtr_rel_year + thtr_rel_month + dvd_rel_year + dvd_rel_month +
```

```
##     imdb_rating + imdb_num_votes + critics_rating + audience_rating +
##     audience_score + best_pic_nom + best_pic_win + best_actor_win +
##     best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - genre            10      1556  74419 3105.1
## - mpaa_rating       5       787  73649 3108.4
## - best_pic_win      1         0  72863 3109.5
## - best_actor_win    1         1  72864 3109.5
## - dvd_rel_month     1         4  72867 3109.5
## - audience_score    1         9  72872 3109.6
## - audience_rating   1        21  72884 3109.7
## - best_pic_nom      1        27  72890 3109.8
## - thtr_rel_month    1        28  72891 3109.8
## - top200_box        1        29  72891 3109.8
## - best_dir_win      1        32  72895 3109.8
## - runtime           1        36  72899 3109.8
## - best_actress_win  1        54  72917 3110.0
## - dvd_rel_year      1       137  72999 3110.7
## <none>                           72863 3111.5
## - title_type        2       735  73598 3114.0
## - thtr_rel_year     1       684  73547 3115.5
## - imdb_num_votes    1       878  73741 3117.2
## - imdb_rating       1     11090  83952 3200.6
## - critics_rating    2    117222 190085 3724.1
##
## Step:  AIC=3105.1
## critics_score ~ title_type + runtime + mpaa_rating + thtr_rel_year +
##     thtr_rel_month + dvd_rel_year + dvd_rel_month + imdb_rating +
##     imdb_num_votes + critics_rating + audience_rating + audience_score +
##     best_pic_nom + best_pic_win + best_actor_win + best_actress_win +
##     best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - mpaa_rating       5       488  74907 3099.3
## - best_pic_win      1         0  74419 3103.1
## - best_actor_win    1         3  74422 3103.1
## - audience_score    1        12  74430 3103.2
## - top200_box        1        19  74437 3103.3
## - dvd_rel_month     1        25  74444 3103.3
## - audience_rating   1        26  74445 3103.3
## - best_dir_win      1        29  74448 3103.3
## - thtr_rel_month    1        38  74457 3103.4
## - best_pic_nom      1        40  74459 3103.4
## - best_actress_win  1        86  74505 3103.8
## - runtime           1        94  74513 3103.9
## - dvd_rel_year      1       124  74542 3104.2
## <none>                           74419 3105.1
```

```
## - title_type        2       557  74975 3105.9
## - thtr_rel_year     1       705  75124 3109.2
## - imdb_num_votes    1      1025  75444 3111.9
## - imdb_rating       1     11877  86296 3198.3
## - critics_rating    2    121234 195653 3722.6
##
## Step:  AIC=3099.3
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
##     critics_rating + audience_rating + audience_score + best_pic_nom +
##     best_pic_win + best_actor_win + best_actress_win + best_dir_win +
##     top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - best_pic_win      1         1  74907 3097.3
## - best_actor_win    1         7  74914 3097.4
## - best_dir_win      1        23  74930 3097.5
## - dvd_rel_month     1        23  74930 3097.5
## - thtr_rel_month    1        26  74933 3097.5
## - audience_score    1        26  74933 3097.5
## - best_pic_nom      1        30  74937 3097.6
## - audience_rating   1        34  74941 3097.6
## - runtime           1        45  74952 3097.7
## - top200_box        1        46  74952 3097.7
## - best_actress_win  1        81  74988 3098.0
## - dvd_rel_year      1       225  75131 3099.2
## <none>                           74907 3099.3
## - title_type        2      1119  76026 3104.8
## - thtr_rel_year     1      1119  76025 3106.8
## - imdb_num_votes    1      1216  76122 3107.7
## - imdb_rating       1     11954  86861 3192.5
## - critics_rating    2    123786 198693 3722.6
##
## Step:  AIC=3097.31
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
##     critics_rating + audience_rating + audience_score + best_pic_nom +
##     best_actor_win + best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - best_actor_win    1         7  74914 3095.4
## - dvd_rel_month     1        24  74931 3095.5
## - audience_score    1        26  74934 3095.5
## - thtr_rel_month    1        27  74934 3095.5
## - best_dir_win      1        27  74935 3095.5
## - audience_rating   1        34  74942 3095.6
## - best_pic_nom      1        41  74948 3095.7
## - runtime           1        45  74952 3095.7
```

```
## - top200_box        1        46  74953 3095.7
## - best_actress_win  1        82  74990 3096.0
## - dvd_rel_year      1       225  75133 3097.2
## <none>                           74907 3097.3
## - title_type        2      1127  76035 3102.9
## - thtr_rel_year     1      1127  76035 3104.9
## - imdb_num_votes    1      1237  76144 3105.8
## - imdb_rating       1     11954  86861 3190.5
## - critics_rating    2    123788 198695 3720.6
##
## Step:  AIC=3095.36
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
##     critics_rating + audience_rating + audience_score + best_pic_nom +
##     best_actress_win + best_dir_win + top200_box
##
##                     Df Sum of Sq    RSS    AIC
## - audience_score     1        26  74940 3093.6
## - dvd_rel_month      1        26  74940 3093.6
## - thtr_rel_month     1        27  74941 3093.6
## - best_dir_win       1        28  74942 3093.6
## - audience_rating    1        36  74950 3093.7
## - best_pic_nom       1        44  74959 3093.7
## - top200_box         1        47  74961 3093.8
## - runtime            1        54  74968 3093.8
## - best_actress_win   1        85  74999 3094.1
## - dvd_rel_year       1       220  75135 3095.3
## <none>                           74914 3095.4
## - title_type         2      1121  76035 3100.9
## - thtr_rel_year      1      1121  76035 3102.9
## - imdb_num_votes     1      1244  76158 3104.0
## - imdb_rating        1     11990  86905 3188.8
## - critics_rating     2    123857 198771 3718.8
##
## Step:  AIC=3093.59
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
##     critics_rating + audience_rating + best_pic_nom + best_actress_win +
##     best_dir_win + top200_box
##
##                     Df Sum of Sq    RSS    AIC
## - audience_rating    1        11  74951 3091.7
## - dvd_rel_month      1        26  74966 3091.8
## - best_dir_win       1        28  74968 3091.8
## - thtr_rel_month     1        30  74970 3091.8
## - top200_box         1        46  74986 3092.0
## - runtime            1        48  74988 3092.0
## - best_pic_nom       1        51  74991 3092.0
```

```
## - best_actress_win  1        79  75019 3092.3
## - dvd_rel_year      1       210  75150 3093.4
## <none>                          74940 3093.6
## - title_type        2      1135  76075 3099.3
## - thtr_rel_year     1      1124  76064 3101.2
## - imdb_num_votes    1      1230  76170 3102.1
## - imdb_rating       1     22117  97057 3257.9
## - critics_rating    2    124565 199505 3719.2
##
## Step:  AIC=3091.68
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
##     critics_rating + best_pic_nom + best_actress_win + best_dir_win +
##     top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - dvd_rel_month     1        26  74977 3089.9
## - thtr_rel_month    1        29  74980 3089.9
## - best_dir_win      1        30  74981 3089.9
## - top200_box        1        45  74996 3090.1
## - runtime           1        49  75000 3090.1
## - best_pic_nom      1        50  75001 3090.1
## - best_actress_win  1        82  75033 3090.4
## - dvd_rel_year      1       221  75172 3091.6
## <none>                          74951 3091.7
## - title_type        2      1127  76078 3097.3
## - thtr_rel_year     1      1127  76079 3099.3
## - imdb_num_votes    1      1228  76179 3100.1
## - imdb_rating       1     29216 104168 3301.3
## - critics_rating    2    127368 202319 3726.2
##
## Step:  AIC=3089.9
## critics_score ~ title_type + runtime + thtr_rel_year + thtr_rel_month +
##     dvd_rel_year + imdb_rating + imdb_num_votes + critics_rating +
##     best_pic_nom + best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - thtr_rel_month    1        21  74998 3088.1
## - best_dir_win      1        34  75011 3088.2
## - top200_box        1        44  75020 3088.3
## - runtime           1        50  75026 3088.3
## - best_pic_nom      1        50  75027 3088.3
## - best_actress_win  1        84  75060 3088.6
## - dvd_rel_year      1       227  75204 3089.8
## <none>                          74977 3089.9
## - title_type        2      1127  76104 3095.5
## - thtr_rel_year     1      1143  76120 3097.6
## - imdb_num_votes    1      1232  76208 3098.4
```

```
## - imdb_rating       1     29235 104212 3299.6
## - critics_rating    2    127435 202412 3724.5
##
## Step:  AIC=3088.08
## critics_score ~ title_type + runtime + thtr_rel_year + dvd_rel_year +
##     imdb_rating + imdb_num_votes + critics_rating + best_pic_nom +
##     best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - best_dir_win      1        33  75031 3086.4
## - runtime           1        40  75038 3086.4
## - top200_box        1        40  75038 3086.4
## - best_pic_nom      1        44  75042 3086.5
## - best_actress_win  1        85  75083 3086.8
## - dvd_rel_year      1       230  75228 3088.1
## <none>                           74998 3088.1
## - title_type        2      1128  76126 3093.7
## - thtr_rel_year     1      1156  76154 3095.9
## - imdb_num_votes    1      1224  76222 3096.5
## - imdb_rating       1     29219 104217 3297.6
## - critics_rating    2    127584 202582 3723.0
##
## Step:  AIC=3086.36
## critics_score ~ title_type + runtime + thtr_rel_year + dvd_rel_year +
##     imdb_rating + imdb_num_votes + critics_rating + best_pic_nom +
##     best_actress_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - top200_box        1        38  75069 3084.7
## - best_pic_nom      1        48  75078 3084.8
## - runtime           1        54  75084 3084.8
## - best_actress_win  1        86  75117 3085.1
## - dvd_rel_year      1       226  75257 3086.3
## <none>                           75031 3086.4
## - title_type        2      1113  76144 3091.8
## - thtr_rel_year     1      1180  76211 3094.4
## - imdb_num_votes    1      1198  76228 3094.5
## - imdb_rating       1     29219 104249 3295.8
## - critics_rating    2    128478 203509 3724.0
##
## Step:  AIC=3084.69
## critics_score ~ title_type + runtime + thtr_rel_year + dvd_rel_year +
##     imdb_rating + imdb_num_votes + critics_rating + best_pic_nom +
##     best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## - best_pic_nom      1        44  75113 3083.1
## - runtime           1        56  75124 3083.2
```

```
## - best_actress_win  1        91  75159 3083.5
## - dvd_rel_year      1       228  75296 3084.6
## <none>                          75069 3084.7
## - title_type        2      1125  76193 3090.3
## - imdb_num_votes    1      1165  76234 3092.6
## - thtr_rel_year     1      1243  76311 3093.2
## - imdb_rating       1     29187 104256 3293.9
## - critics_rating    2    129010 204078 3723.8
##
## Step:  AIC=3083.07
## critics_score ~ title_type + runtime + thtr_rel_year + dvd_rel_year +
##     imdb_rating + imdb_num_votes + critics_rating + best_actress_win
##
##                     Df Sum of Sq    RSS    AIC
## - runtime            1        70  75182 3081.7
## - best_actress_win   1       112  75225 3082.0
## <none>                           75113 3083.1
## - dvd_rel_year       1       239  75352 3083.1
## - title_type         2      1115  76228 3088.5
## - imdb_num_votes     1      1121  76234 3090.6
## - thtr_rel_year      1      1298  76410 3092.1
## - imdb_rating        1     29337 104449 3293.1
## - critics_rating     2    129881 204994 3724.6
##
## Step:  AIC=3081.66
## critics_score ~ title_type + thtr_rel_year + dvd_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating + best_actress_win
##
##                     Df Sum of Sq    RSS    AIC
## - best_actress_win   1       146  75329 3080.9
## <none>                           75182 3081.7
## - dvd_rel_year       1       246  75429 3081.8
## - title_type         2      1067  76250 3086.7
## - imdb_num_votes     1      1052  76234 3088.6
## - thtr_rel_year      1      1389  76571 3091.4
## - imdb_rating        1     31182 106364 3302.8
## - critics_rating     2    130168 205351 3723.7
##
## Step:  AIC=3080.91
## critics_score ~ title_type + thtr_rel_year + dvd_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating
##
##                   Df Sum of Sq    RSS    AIC
## - dvd_rel_year     1       230  75559 3080.9
## <none>                          75329 3080.9
## - title_type       2      1028  76357 3085.6
## - imdb_num_votes   1      1008  76337 3087.5
## - thtr_rel_year    1      1400  76729 3090.8
```

```
## - imdb_rating     1     31451 106780 3303.3
## - critics_rating  2    130185 205514 3722.3
##
## Step:  AIC=3080.88
## critics_score ~ title_type + thtr_rel_year + imdb_rating + imdb_num_votes
+
##     critics_rating
##
##                   Df Sum of Sq    RSS    AIC
## <none>                          75559 3080.9
## - title_type      2      1124  76683 3086.4
## - imdb_num_votes  1       972  76532 3087.1
## - thtr_rel_year   1      1300  76860 3089.8
## - imdb_rating     1     31222 106782 3301.3
## - critics_rating  2    133268 208827 3730.5
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + thtr_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.714  -7.614   0.008   7.998  29.240
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.055e+02  8.547e+01   3.574 0.000378 ***
## title_typeFeature Film -5.105e+00  1.780e+00  -2.867 0.004278 **
## title_typeTV Movie       8.250e-01  5.171e+00   0.160 0.873290
## thtr_rel_year           -1.396e-01  4.224e-02  -3.306 0.001001 **
## imdb_rating              9.081e+00  5.606e-01  16.198  < 2e-16 ***
## imdb_num_votes          -1.302e-05  4.554e-06  -2.859 0.004395 **
## critics_ratingFresh     -8.151e+00  1.336e+00  -6.101 1.83e-09 ***
## critics_ratingRotten    -4.195e+01  1.441e+00 -29.119  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 635 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8519
## F-statistic: 528.5 on 7 and 635 DF,  p-value: < 2.2e-16
```

```
# checking vif for the variables in the final model..
vif(data.frame(mod_data[,c("thtr_rel_year","imdb_rating","imdb_num_votes")]))
```

```
##        Variables      VIF
## 1  thtr_rel_year 1.033101
```

```
## 2    imdb_rating 1.135759
## 3 imdb_num_votes 1.160391
```

Our vif is in acceptable range. So using step function our initial model with 20 variables gets reduced to only 5 variables.

Now we will use backward elimination to find the final equation..

First of all, I am going to add all the variables in the model and use backward elimination to make the final model. I am going to do backward elimination by 'Adjusted R squared technique' as I think this method gives good robust results and also I look at p- values of variables in the model to do backward elimination, the significance level will be 0.05 for p-value.

```
#assumed model

m1 <- lm(critics_score ~. , data= mod_data)

summary(m1)

##
## Call:
## lm(formula = critics_score ~ ., data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.498  -7.261  -0.111   7.318  29.747
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.288e+01  2.112e+02   0.203  0.83921
## title_typeFeature Film         -9.605e+00  4.103e+00  -2.341  0.01955 *
## title_typeTV Movie             -5.303e+00  6.457e+00  -0.821  0.41185
## genreAnimation                 -6.408e-01  4.397e+00  -0.146  0.88417
## genreArt House & International -6.443e+00  3.499e+00  -1.842  0.06600 .
## genreComedy                     1.093e+00  1.872e+00   0.584  0.55967
## genreDocumentary               -5.721e+00  4.416e+00  -1.296  0.19558
## genreDrama                      2.335e+00  1.657e+00   1.409  0.15947
## genreHorror                     2.598e+00  2.781e+00   0.934  0.35064
## genreMusical & Performing Arts  1.071e+00  3.803e+00   0.282  0.77834
## genreMystery & Suspense         1.203e+00  2.106e+00   0.571  0.56804
## genreOther                      1.166e+00  3.165e+00   0.368  0.71279
## genreScience Fiction & Fantasy -9.493e-01  4.146e+00  -0.229  0.81896
## runtime                         1.547e-02  2.821e-02   0.548  0.58364
## mpaa_ratingNC-17               -1.682e+00  8.405e+00  -0.200  0.84143
## mpaa_ratingPG                  -2.836e+00  3.134e+00  -0.905  0.36589
## mpaa_ratingPG-13               -5.110e+00  3.280e+00  -1.558  0.11972
## mpaa_ratingR                   -4.560e+00  3.154e+00  -1.446  0.14872
## mpaa_ratingUnrated             -6.698e-01  3.680e+00  -0.182  0.85564
```

```
## thtr_rel_year                   -1.461e-01  6.116e-02  -2.389  0.01722 *
## thtr_rel_month                  -6.259e-02  1.300e-01  -0.482  0.63031
## dvd_rel_year                     1.404e-01  1.315e-01   1.067  0.28618
## dvd_rel_month                   -2.383e-02  1.331e-01  -0.179  0.85793
## imdb_rating                      8.923e+00  9.276e-01   9.620  < 2e-16 ***
## imdb_num_votes                  -1.431e-05  5.287e-06  -2.707  0.00699 **
## critics_ratingFresh             -7.967e+00  1.390e+00  -5.733 1.55e-08 ***
## critics_ratingRotten            -4.121e+01  1.519e+00 -27.124  < 2e-16 ***
## audience_ratingUpright          -7.592e-01  1.807e+00  -0.420  0.67457
## audience_score                   1.770e-02  6.449e-02   0.274  0.78389
## best_pic_nomyes                  1.402e+00  2.926e+00   0.479  0.63213
## best_pic_winyes                  1.382e-02  5.119e+00   0.003  0.99785
## best_actor_winyes                1.050e-01  1.322e+00   0.079  0.93669
## best_actress_winyes              9.754e-01  1.455e+00   0.670  0.50295
## best_dir_winyes                  9.826e-01  1.906e+00   0.516  0.60631
## top200_boxyes                    1.517e+00  3.104e+00   0.489  0.62527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.95 on 608 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8508
## F-statistic: 108.7 on 34 and 608 DF,  p-value: < 2.2e-16
```

Now comparing p-value of all variables in the model, we see that "best_pic_win" variable has the highest p-value, so we will remove the variable.

```
# without best_pic_win

m2 <- lm(critics_score ~. -best_pic_win, data= mod_data)

summary(m2)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.498  -7.262  -0.111   7.318  29.747
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.288e+01  2.111e+02   0.203  0.83905
## title_typeFeature Film         -9.605e+00  4.099e+00  -2.343  0.01944 *
## title_typeTV Movie             -5.302e+00  6.452e+00  -0.822  0.41147
## genreAnimation                 -6.407e-01  4.393e+00  -0.146  0.88409
## genreArt House & International -6.443e+00  3.494e+00  -1.844  0.06569 .
## genreComedy                     1.093e+00  1.869e+00   0.585  0.55883
```

```
## genreDocumentary                -5.721e+00  4.411e+00  -1.297  0.19510
## genreDrama                       2.335e+00  1.655e+00   1.410  0.15894
## genreHorror                      2.598e+00  2.779e+00   0.935  0.35022
## genreMusical & Performing Arts   1.071e+00  3.800e+00   0.282  0.77810
## genreMystery & Suspense          1.203e+00  2.104e+00   0.572  0.56758
## genreOther                       1.165e+00  3.156e+00   0.369  0.71212
## genreScience Fiction & Fantasy -9.494e-01  4.142e+00  -0.229  0.81876
## runtime                          1.547e-02  2.819e-02   0.549  0.58326
## mpaa_ratingNC-17                -1.682e+00  8.398e+00  -0.200  0.84131
## mpaa_ratingPG                   -2.836e+00  3.131e+00  -0.906  0.36550
## mpaa_ratingPG-13                -5.110e+00  3.276e+00  -1.560  0.11927
## mpaa_ratingR                    -4.561e+00  3.151e+00  -1.447  0.14836
## mpaa_ratingUnrated              -6.697e-01  3.677e+00  -0.182  0.85553
## thtr_rel_year                   -1.461e-01  6.101e-02  -2.395  0.01693 *
## thtr_rel_month                  -6.262e-02  1.296e-01  -0.483  0.62916
## dvd_rel_year                     1.404e-01  1.314e-01   1.069  0.28566
## dvd_rel_month                   -2.386e-02  1.327e-01  -0.180  0.85738
## imdb_rating                      8.923e+00  9.269e-01   9.628  < 2e-16 ***
## imdb_num_votes                  -1.431e-05  5.205e-06  -2.749  0.00616 **
## critics_ratingFresh             -7.967e+00  1.387e+00  -5.746 1.45e-08 ***
## critics_ratingRotten            -4.121e+01  1.518e+00 -27.151  < 2e-16 ***
## audience_ratingUpright          -7.591e-01  1.806e+00  -0.420  0.67430
## audience_score                   1.769e-02  6.439e-02   0.275  0.78364
## best_pic_nomyes                  1.405e+00  2.683e+00   0.524  0.60069
## best_actor_winyes                1.048e-01  1.317e+00   0.080  0.93662
## best_actress_winyes              9.756e-01  1.453e+00   0.672  0.50211
## best_dir_winyes                  9.840e-01  1.832e+00   0.537  0.59145
## top200_boxyes                    1.517e+00  3.101e+00   0.489  0.62499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.94 on 609 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8511
## F-statistic: 112.2 on 33 and 609 DF,  p-value: < 2.2e-16
```

```
# checking vif for the numeric variables..
mod_data2 <- select_if(mod_data, is.numeric)

vif(data.frame(mod_data2[,-8]))
```

```
##        Variables      VIF
## 1        runtime 1.273205
## 2  thtr_rel_year 1.850518
## 3 thtr_rel_month 1.089128
## 4   dvd_rel_year 1.792483
## 5  dvd_rel_month 1.040576
## 6    imdb_rating 4.217520
```

```
## 7 imdb_num_votes 1.289179
## 8 audience_score 4.018047
```

We see that in our model now the adjusted R squared value is increased a bit, which is what we want, now looking at the model again we find that “best_actor_win” has the highest p-value now, so we will remove the variable.

```
# without audience_score..

m3 <- lm(critics_score ~. -best_pic_win -best_actor_win, data= mod_data)

summary(m3)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win,
##     data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.508  -7.267  -0.111   7.310  29.738
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.402e+01  2.104e+02   0.209  0.83437
## title_typeFeature Film         -9.600e+00  4.095e+00  -2.344  0.01939 *
## title_typeTV Movie             -5.311e+00  6.445e+00  -0.824  0.41026
## genreAnimation                 -6.322e-01  4.388e+00  -0.144  0.88548
## genreArt House & International -6.450e+00  3.490e+00  -1.848  0.06510 .
## genreComedy                     1.093e+00  1.867e+00   0.585  0.55866
## genreDocumentary               -5.718e+00  4.407e+00  -1.297  0.19496
## genreDrama                      2.338e+00  1.653e+00   1.414  0.15786
## genreHorror                     2.593e+00  2.776e+00   0.934  0.35058
## genreMusical & Performing Arts  1.066e+00  3.796e+00   0.281  0.77887
## genreMystery & Suspense         1.216e+00  2.096e+00   0.580  0.56189
## genreOther                      1.169e+00  3.153e+00   0.371  0.71098
## genreScience Fiction & Fantasy -9.581e-01  4.137e+00  -0.232  0.81692
## runtime                         1.587e-02  2.772e-02   0.572  0.56722
## mpaa_ratingNC-17               -1.635e+00  8.370e+00  -0.195  0.84519
## mpaa_ratingPG                  -2.827e+00  3.127e+00  -0.904  0.36624
## mpaa_ratingPG-13               -5.108e+00  3.273e+00  -1.561  0.11910
## mpaa_ratingR                   -4.559e+00  3.149e+00  -1.448  0.14813
## mpaa_ratingUnrated             -6.709e-01  3.674e+00  -0.183  0.85516
## thtr_rel_year                  -1.458e-01  6.083e-02  -2.397  0.01685 *
## thtr_rel_month                 -6.249e-02  1.295e-01  -0.483  0.62956
## dvd_rel_year                    1.395e-01  1.308e-01   1.067  0.28657
## dvd_rel_month                  -2.482e-02  1.320e-01  -0.188  0.85093
## imdb_rating                     8.925e+00  9.258e-01   9.641  < 2e-16 ***
```

```
## imdb_num_votes                  -1.432e-05  5.198e-06  -2.756  0.00603 **
## critics_ratingFresh             -7.963e+00  1.385e+00  -5.751  1.4e-08 ***
## critics_ratingRotten            -4.121e+01  1.516e+00 -27.174  < 2e-16 ***
## audience_ratingUpright          -7.657e-01  1.802e+00  -0.425  0.67109
## audience_score                   1.769e-02  6.434e-02   0.275  0.78349
## best_pic_nomyes                  1.426e+00  2.667e+00   0.535  0.59310
## best_actress_winyes              9.812e-01  1.450e+00   0.677  0.49884
## best_dir_winyes                  9.864e-01  1.831e+00   0.539  0.59021
## top200_boxyes                    1.524e+00  3.097e+00   0.492  0.62288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 610 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8513
## F-statistic: 115.9 on 32 and 610 DF,  p-value: < 2.2e-16
```

Looking again at the model we will remove "audience_score" variable as it's p-value is highest and also the vif of audience_score is high, so there will be multicollinearity problems.

```
# audience_score variable..

m4 <- lm(formula = critics_score ~ . - best_pic_win -
best_actor_win-audience_score,
    data = mod_data)

summary(m4)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.457  -7.208  -0.063   7.371  29.693
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.960e+01  2.093e+02   0.237  0.81270
## title_typeFeature Film         -9.563e+00  4.090e+00  -2.338  0.01970 *
## title_typeTV Movie             -5.307e+00  6.441e+00  -0.824  0.41025
## genreAnimation                 -5.755e-01  4.380e+00  -0.131  0.89550
## genreArt House & International -6.488e+00  3.485e+00  -1.862  0.06314 .
## genreComedy                     1.122e+00  1.863e+00   0.602  0.54724
## genreDocumentary               -5.674e+00  4.401e+00  -1.289  0.19775
## genreDrama                      2.334e+00  1.652e+00   1.412  0.15834
## genreHorror                     2.566e+00  2.772e+00   0.926  0.35500
```

```
## genreMusical & Performing Arts  1.140e+00  3.784e+00   0.301  0.76331
## genreMystery & Suspense         1.168e+00  2.087e+00   0.560  0.57579
## genreOther                      1.165e+00  3.151e+00   0.370  0.71173
## genreScience Fiction & Fantasy -9.616e-01  4.134e+00  -0.233  0.81613
## runtime                         1.542e-02  2.765e-02   0.557  0.57739
## mpaa_ratingNC-17               -1.656e+00  8.364e+00  -0.198  0.84314
## mpaa_ratingPG                  -2.826e+00  3.124e+00  -0.905  0.36602
## mpaa_ratingPG-13               -5.123e+00  3.270e+00  -1.567  0.11772
## mpaa_ratingR                   -4.579e+00  3.146e+00  -1.456  0.14602
## mpaa_ratingUnrated             -6.656e-01  3.671e+00  -0.181  0.85618
## thtr_rel_year                  -1.457e-01  6.078e-02  -2.398  0.01680 *
## thtr_rel_month                 -6.500e-02  1.291e-01  -0.504  0.61471
## dvd_rel_year                    1.366e-01  1.302e-01   1.049  0.29479
## dvd_rel_month                  -2.459e-02  1.319e-01  -0.186  0.85221
## imdb_rating                     9.092e+00  6.978e-01  13.031  < 2e-16 ***
## imdb_num_votes                 -1.425e-05  5.187e-06  -2.747  0.00619 **
## critics_ratingFresh            -7.966e+00  1.383e+00  -5.758 1.35e-08 ***
## critics_ratingRotten           -4.123e+01  1.513e+00 -27.251  < 2e-16 ***
## audience_ratingUpright         -4.158e-01  1.275e+00  -0.326  0.74444
## best_pic_nomyes                 1.488e+00  2.656e+00   0.560  0.57544
## best_actress_winyes             9.557e-01  1.446e+00   0.661  0.50884
## best_dir_winyes                 9.848e-01  1.829e+00   0.538  0.59051
## top200_boxyes                   1.511e+00  3.095e+00   0.488  0.62555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.92 on 611 degrees of freedom
## Multiple R-squared:  0.8587, Adjusted R-squared:  0.8515
## F-statistic: 119.8 on 31 and 611 DF,  p-value: < 2.2e-16
```

```
# checking vif..
mod_data2 <- select_if(mod_data, is.numeric)
vif(data.frame(mod_data2[,c(-8,-9)]))
```

```
##        Variables      VIF
## 1        runtime 1.256706
## 2  thtr_rel_year 1.850510
## 3 thtr_rel_month 1.087338
## 4   dvd_rel_year 1.778496
## 5  dvd_rel_month 1.040423
## 6    imdb_rating 1.171291
## 7 imdb_num_votes 1.285317
```

Now we can see that vif are all within considerable range after removing audience_score variable..

We will remove “runtime” variable as it’s p-value is highest.

```r
# removing runtime variable

m5 <- lm(formula = critics_score ~ . - best_pic_win -
best_actor_win-audience_score-runtime,
    data = mod_data)

summary(m5)
```

```
##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.082  -7.261  -0.100   7.395  29.798
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.665e+01  2.088e+02   0.271   0.7862
## title_typeFeature Film         -9.560e+00  4.088e+00  -2.339   0.0197 *
## title_typeTV Movie             -5.354e+00  6.436e+00  -0.832   0.4059
## genreAnimation                 -6.313e-01  4.376e+00  -0.144   0.8853
## genreArt House & International -6.513e+00  3.483e+00  -1.870   0.0620 .
## genreComedy                     1.019e+00  1.853e+00   0.550   0.5824
## genreDocumentary               -5.805e+00  4.392e+00  -1.322   0.1868
## genreDrama                      2.397e+00  1.647e+00   1.455   0.1461
## genreHorror                     2.403e+00  2.755e+00   0.872   0.3834
## genreMusical & Performing Arts  1.269e+00  3.774e+00   0.336   0.7368
## genreMystery & Suspense         1.224e+00  2.083e+00   0.588   0.5570
## genreOther                      1.195e+00  3.148e+00   0.380   0.7043
## genreScience Fiction & Fantasy -1.001e+00  4.131e+00  -0.242   0.8086
## mpaa_ratingNC-17               -1.587e+00  8.358e+00  -0.190   0.8494
## mpaa_ratingPG                  -2.670e+00  3.110e+00  -0.858   0.3910
## mpaa_ratingPG-13               -4.831e+00  3.226e+00  -1.498   0.1348
## mpaa_ratingR                   -4.376e+00  3.123e+00  -1.401   0.1616
## mpaa_ratingUnrated             -3.395e-01  3.622e+00  -0.094   0.9254
## thtr_rel_year                  -1.512e-01  5.994e-02  -2.523   0.0119 *
## thtr_rel_month                 -5.092e-02  1.265e-01  -0.403   0.6874
## dvd_rel_year                    1.390e-01  1.301e-01   1.068   0.2858
## dvd_rel_month                  -2.379e-02  1.318e-01  -0.180   0.8569
## imdb_rating                     9.142e+00  6.917e-01  13.218  < 2e-16 ***
## imdb_num_votes                 -1.365e-05  5.072e-06  -2.692   0.0073 **
## critics_ratingFresh            -7.950e+00  1.382e+00  -5.751  1.4e-08 ***
## critics_ratingRotten           -4.116e+01  1.507e+00 -27.306  < 2e-16 ***
## audience_ratingUpright         -4.442e-01  1.273e+00  -0.349   0.7273
## best_pic_nomyes                 1.607e+00  2.645e+00   0.608   0.5437
## best_actress_winyes             1.046e+00  1.436e+00   0.728   0.4666
```

```
## best_dir_winyes                 1.135e+00  1.808e+00   0.627   0.5306
## top200_boxyes                   1.574e+00  3.091e+00   0.509   0.6107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 612 degrees of freedom
## Multiple R-squared:  0.8586, Adjusted R-squared:  0.8517
## F-statistic: 123.9 on 30 and 612 DF,  p-value: < 2.2e-16
```

Now we will remove "mpaa_rating" variable as p-value is very high.

```
# without genre

m6 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime - mpaa_rating, data = mod_data)

summary(m6)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.210  -7.481  -0.019   7.423  29.674
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.273e+00  2.060e+02   0.026  0.97959
## title_typeFeature Film         -1.051e+01  4.062e+00  -2.588  0.00989 **
## title_typeTV Movie             -5.036e+00  6.432e+00  -0.783  0.43397
## genreAnimation                  2.060e+00  3.956e+00   0.521  0.60269
## genreArt House & International -6.258e+00  3.423e+00  -1.828  0.06803 .
## genreComedy                     5.764e-01  1.841e+00   0.313  0.75434
## genreDocumentary               -5.106e+00  4.354e+00  -1.173  0.24135
## genreDrama                      1.790e+00  1.611e+00   1.111  0.26708
## genreHorror                     2.006e+00  2.694e+00   0.745  0.45683
## genreMusical & Performing Arts  8.911e-01  3.766e+00   0.237  0.81301
## genreMystery & Suspense         5.219e-01  2.040e+00   0.256  0.79820
## genreOther                      1.075e+00  3.138e+00   0.342  0.73211
## genreScience Fiction & Fantasy -1.036e+00  4.130e+00  -0.251  0.80198
## thtr_rel_year                  -1.763e-01  5.683e-02  -3.102  0.00201 **
## thtr_rel_month                 -5.117e-02  1.262e-01  -0.405  0.68538
## dvd_rel_year                    1.881e-01  1.283e-01   1.467  0.14300
## dvd_rel_month                  -2.821e-02  1.317e-01  -0.214  0.83048
## imdb_rating                     9.258e+00  6.901e-01  13.415  < 2e-16 ***
## imdb_num_votes                 -1.561e-05  5.015e-06  -3.112  0.00194 **
```

```
## critics_ratingFresh            -8.017e+00  1.378e+00  -5.817 9.64e-09 ***
## critics_ratingRotten           -4.140e+01  1.497e+00 -27.664  < 2e-16 ***
## audience_ratingUpright         -4.126e-01  1.272e+00  -0.324  0.74586
## best_pic_nomyes                 1.563e+00  2.644e+00   0.591  0.55479
## best_actress_winyes             1.038e+00  1.434e+00   0.724  0.46937
## best_dir_winyes                 9.956e-01  1.808e+00   0.551  0.58201
## top200_boxyes                   2.375e+00  3.059e+00   0.776  0.43788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 617 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.8513
## F-statistic: 148.1 on 25 and 617 DF,  p-value: < 2.2e-16
```

But we see that even though we removed mpaa_rating variable our Adjusted R Squared value is reduced, but it is reduced by a very small amount by 0.0004 so it doesn't affect the model very much. Now we remove "audience_rating" variable.

```
# without audience_rating..

m7 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime - mpaa_rating-audience_rating, data = mod_data)

summary(m7)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating,
##     data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.262  -7.525   0.023   7.385  29.782
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.718e+00  2.047e+02  -0.008  0.99331
## title_typeFeature Film         -1.048e+01  4.058e+00  -2.583  0.01003 *
## title_typeTV Movie             -5.058e+00  6.427e+00  -0.787  0.43158
## genreAnimation                  1.936e+00  3.934e+00   0.492  0.62283
## genreArt House & International -6.331e+00  3.413e+00  -1.855  0.06406 .
## genreComedy                     5.557e-01  1.839e+00   0.302  0.76257
## genreDocumentary               -5.125e+00  4.350e+00  -1.178  0.23921
## genreDrama                      1.762e+00  1.608e+00   1.096  0.27353
## genreHorror                     2.059e+00  2.688e+00   0.766  0.44388
## genreMusical & Performing Arts  8.410e-01  3.760e+00   0.224  0.82307
## genreMystery & Suspense         5.646e-01  2.035e+00   0.278  0.78149
```

```
## genreOther                         1.042e+00  3.134e+00   0.333  0.73959
## genreScience Fiction & Fantasy -1.005e+00  4.126e+00  -0.244  0.80770
## thtr_rel_year                    -1.765e-01  5.679e-02  -3.107  0.00198 **
## thtr_rel_month                   -4.991e-02  1.261e-01  -0.396  0.69237
## dvd_rel_year                      1.919e-01  1.276e-01   1.504  0.13304
## dvd_rel_month                    -2.763e-02  1.316e-01  -0.210  0.83373
## imdb_rating                       9.148e+00  6.021e-01  15.194  < 2e-16 ***
## imdb_num_votes                   -1.562e-05  5.011e-06  -3.117  0.00191 **
## critics_ratingFresh              -7.971e+00  1.370e+00  -5.818 9.54e-09 ***
## critics_ratingRotten             -4.132e+01  1.474e+00 -28.037  < 2e-16 ***
## best_pic_nomyes                   1.554e+00  2.642e+00   0.588  0.55669
## best_actress_winyes               1.061e+00  1.431e+00   0.741  0.45876
## best_dir_winyes                   1.030e+00  1.803e+00   0.571  0.56822
## top200_boxyes                     2.345e+00  3.056e+00   0.767  0.44311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.92 on 618 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.8515
## F-statistic: 154.4 on 24 and 618 DF,  p-value: < 2.2e-16
```

Now we will remove “dvd_rel_month” variable.

```
# without dvd_rel_month variable.

m9 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime - mpaa_rating-audience_rating-dvd_rel_month,
data = mod_data)

summary(m9)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.312  -7.548   0.019   7.314  29.812
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.116e+00  2.045e+02  -0.015  0.98785
## title_typeFeature Film         -1.052e+01  4.051e+00  -2.597  0.00964 **
## title_typeTV Movie             -5.116e+00  6.416e+00  -0.797  0.42556
## genreAnimation                  1.978e+00  3.926e+00   0.504  0.61451
## genreArt House & International -6.336e+00  3.410e+00  -1.858  0.06365 .
```

```
## genreComedy                       5.725e-01  1.836e+00   0.312  0.75521
## genreDocumentary                 -5.148e+00  4.345e+00  -1.185  0.23655
## genreDrama                        1.782e+00  1.604e+00   1.111  0.26688
## genreHorror                       2.057e+00  2.685e+00   0.766  0.44409
## genreMusical & Performing Arts    8.695e-01  3.754e+00   0.232  0.81692
## genreMystery & Suspense           5.819e-01  2.031e+00   0.286  0.77464
## genreOther                        1.068e+00  3.129e+00   0.341  0.73291
## genreScience Fiction & Fantasy   -1.038e+00  4.120e+00  -0.252  0.80126
## thtr_rel_year                    -1.770e-01  5.669e-02  -3.123  0.00188 **
## thtr_rel_month                   -4.544e-02  1.242e-01  -0.366  0.71456
## dvd_rel_year                      1.931e-01  1.274e-01   1.516  0.13001
## imdb_rating                       9.139e+00  6.000e-01  15.231  < 2e-16 ***
## imdb_num_votes                   -1.562e-05  5.007e-06  -3.120  0.00189 **
## critics_ratingFresh              -7.959e+00  1.368e+00  -5.819 9.49e-09 ***
## critics_ratingRotten             -4.131e+01  1.472e+00 -28.059  < 2e-16 ***
## best_pic_nomyes                   1.555e+00  2.640e+00   0.589  0.55609
## best_actress_winyes               1.063e+00  1.430e+00   0.743  0.45759
## best_dir_winyes                   1.055e+00  1.798e+00   0.587  0.55740
## top200_boxyes                     2.338e+00  3.053e+00   0.766  0.44404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 619 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.8518
## F-statistic: 161.4 on 23 and 619 DF,  p-value: < 2.2e-16
```

Now we will remove “thtr_rel_month”..

```
# Without thtr_rel_month variable

m10 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month, data = mod_data)

summary(m10)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.393  -7.538   0.011   7.455  30.078
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    -3.713e+00  2.043e+02  -0.018  0.98551
## title_typeFeature Film         -1.051e+01  4.048e+00  -2.596  0.00965 **
## title_typeTV Movie             -5.039e+00  6.408e+00  -0.786  0.43199
## genreAnimation                  1.937e+00  3.922e+00   0.494  0.62150
## genreArt House & International -6.350e+00  3.408e+00  -1.863  0.06287 .
## genreComedy                     5.547e-01  1.834e+00   0.302  0.76239
## genreDocumentary               -5.132e+00  4.342e+00  -1.182  0.23765
## genreDrama                      1.778e+00  1.603e+00   1.110  0.26757
## genreHorror                     2.056e+00  2.684e+00   0.766  0.44397
## genreMusical & Performing Arts  8.301e-01  3.750e+00   0.221  0.82490
## genreMystery & Suspense         6.053e-01  2.029e+00   0.298  0.76556
## genreOther                      1.111e+00  3.125e+00   0.355  0.72244
## genreScience Fiction & Fantasy -1.001e+00  4.116e+00  -0.243  0.80792
## thtr_rel_year                  -1.773e-01  5.664e-02  -3.131  0.00183 **
## dvd_rel_year                    1.936e-01  1.273e-01   1.521  0.12874
## imdb_rating                     9.129e+00  5.989e-01  15.242  < 2e-16 ***
## imdb_num_votes                 -1.566e-05  5.003e-06  -3.130  0.00183 **
## critics_ratingFresh            -7.963e+00  1.367e+00  -5.826 9.11e-09 ***
## critics_ratingRotten           -4.133e+01  1.470e+00 -28.107  < 2e-16 ***
## best_pic_nomyes                 1.426e+00  2.615e+00   0.545  0.58576
## best_actress_winyes             1.056e+00  1.429e+00   0.739  0.45997
## best_dir_winyes                 1.023e+00  1.794e+00   0.570  0.56892
## top200_boxyes                   2.268e+00  3.045e+00   0.745  0.45676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 620 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.852
## F-statistic:   169 on 22 and 620 DF,  p-value: < 2.2e-16
```

Removing “best_pic_nom” variable..

```
# removing best_pic_nom variable

m11 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom, data =
mod_data)

summary(m11)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month - best_pic_nom, data = mod_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -34.361  -7.546   0.057   7.378  30.120
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -5.028e+00  2.042e+02  -0.025  0.98036
## title_typeFeature Film         -1.050e+01  4.046e+00  -2.595  0.00968 **
## title_typeTV Movie             -5.046e+00  6.405e+00  -0.788  0.43106
## genreAnimation                  1.933e+00  3.920e+00   0.493  0.62216
## genreArt House & International -6.340e+00  3.406e+00  -1.862  0.06312 .
## genreComedy                     5.862e-01  1.832e+00   0.320  0.74904
## genreDocumentary               -5.113e+00  4.339e+00  -1.178  0.23912
## genreDrama                      1.828e+00  1.599e+00   1.143  0.25353
## genreHorror                     2.080e+00  2.682e+00   0.776  0.43825
## genreMusical & Performing Arts  8.234e-01  3.748e+00   0.220  0.82619
## genreMystery & Suspense         6.257e-01  2.027e+00   0.309  0.75773
## genreOther                      1.217e+00  3.117e+00   0.391  0.69630
## genreScience Fiction & Fantasy -1.032e+00  4.113e+00  -0.251  0.80193
## thtr_rel_year                  -1.805e-01  5.631e-02  -3.205  0.00142 **
## dvd_rel_year                    1.974e-01  1.270e-01   1.554  0.12076
## imdb_rating                     9.146e+00  5.978e-01  15.298  < 2e-16 ***
## imdb_num_votes                 -1.512e-05  4.900e-06  -3.085  0.00213 **
## critics_ratingFresh            -8.052e+00  1.356e+00  -5.938  4.8e-09 ***
## critics_ratingRotten           -4.141e+01  1.463e+00 -28.308  < 2e-16 ***
## best_actress_winyes             1.172e+00  1.412e+00   0.830  0.40664
## best_dir_winyes                 1.093e+00  1.789e+00   0.611  0.54128
## top200_boxyes                   2.215e+00  3.042e+00   0.728  0.46673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 621 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8522
## F-statistic: 177.2 on 21 and 621 DF,  p-value: < 2.2e-16
```

Now we will remove “best_dir_win”.

```
# removing thtr_rel_month
m12 <-lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom-best_di
r_win, data = mod_data)

summary(m12)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
```

```
##     dvd_rel_month - thtr_rel_month - best_pic_nom - best_dir_win,
##     data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.332  -7.576  -0.027   7.380  30.109
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.220e+00  2.037e+02   0.011  0.99131
## title_typeFeature Film         -1.043e+01  4.042e+00  -2.580  0.01011 *
## title_typeTV Movie             -5.021e+00  6.401e+00  -0.784  0.43316
## genreAnimation                  1.877e+00  3.917e+00   0.479  0.63186
## genreArt House & International -6.399e+00  3.403e+00  -1.881  0.06049 .
## genreComedy                     5.820e-01  1.831e+00   0.318  0.75067
## genreDocumentary               -5.117e+00  4.337e+00  -1.180  0.23850
## genreDrama                      1.825e+00  1.598e+00   1.142  0.25392
## genreHorror                     2.073e+00  2.680e+00   0.773  0.43965
## genreMusical & Performing Arts  8.555e-01  3.746e+00   0.228  0.81942
## genreMystery & Suspense         6.440e-01  2.026e+00   0.318  0.75072
## genreOther                      1.151e+00  3.114e+00   0.370  0.71170
## genreScience Fiction & Fantasy -9.976e-01  4.111e+00  -0.243  0.80834
## thtr_rel_year                  -1.832e-01  5.611e-02  -3.266  0.00115 **
## dvd_rel_year                    1.965e-01  1.270e-01   1.547  0.12227
## imdb_rating                     9.158e+00  5.972e-01  15.334  < 2e-16 ***
## imdb_num_votes                 -1.469e-05  4.849e-06  -3.031  0.00254 **
## critics_ratingFresh            -8.045e+00  1.355e+00  -5.936 4.85e-09 ***
## critics_ratingRotten           -4.145e+01  1.461e+00 -28.374  < 2e-16 ***
## best_actress_winyes             1.214e+00  1.410e+00   0.861  0.38944
## top200_boxyes                   2.152e+00  3.039e+00   0.708  0.47911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 622 degrees of freedom
## Multiple R-squared:  0.8569, Adjusted R-squared:  0.8523
## F-statistic: 186.2 on 20 and 622 DF,  p-value: < 2.2e-16
```

Now Removing “top200_box” variable.

```
m13 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom-best_di
r_win-top200_box, data = mod_data)

summary(m13)

##
## Call:
```

```
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month - best_pic_nom - best_dir_win -
##     top200_box, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.279  -7.601   0.015   7.376  30.135
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.545e+00  2.034e+02   0.042  0.96651
## title_typeFeature Film         -1.044e+01  4.041e+00  -2.585  0.00997 **
## title_typeTV Movie             -5.020e+00  6.399e+00  -0.784  0.43307
## genreAnimation                  1.724e+00  3.909e+00   0.441  0.65940
## genreArt House & International -6.509e+00  3.398e+00  -1.916  0.05585 .
## genreComedy                     4.619e-01  1.822e+00   0.253  0.79997
## genreDocumentary               -5.220e+00  4.333e+00  -1.205  0.22875
## genreDrama                      1.702e+00  1.588e+00   1.071  0.28438
## genreHorror                     1.935e+00  2.672e+00   0.724  0.46924
## genreMusical & Performing Arts  7.254e-01  3.740e+00   0.194  0.84626
## genreMystery & Suspense         5.112e-01  2.017e+00   0.253  0.79997
## genreOther                      1.046e+00  3.109e+00   0.336  0.73667
## genreScience Fiction & Fantasy -9.396e-01  4.108e+00  -0.229  0.81918
## thtr_rel_year                  -1.872e-01  5.579e-02  -3.356  0.00084 ***
## dvd_rel_year                    1.974e-01  1.269e-01   1.556  0.12024
## imdb_rating                     9.144e+00  5.967e-01  15.325  < 2e-16 ***
## imdb_num_votes                 -1.391e-05  4.719e-06  -2.948  0.00332 **
## critics_ratingFresh            -8.120e+00  1.351e+00  -6.012 3.12e-09 ***
## critics_ratingRotten           -4.153e+01  1.455e+00 -28.535  < 2e-16 ***
## best_actress_winyes             1.266e+00  1.407e+00   0.900  0.36864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 623 degrees of freedom
## Multiple R-squared:  0.8568, Adjusted R-squared:  0.8524
## F-statistic: 196.2 on 19 and 623 DF,  p-value: < 2.2e-16
```

Now removing “best_actress_win” variable.

```
m14 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom-best_di
r_win-top200_box-best_actress_win, data = mod_data)

summary(m14)
```

```
##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month - best_pic_nom - best_dir_win -
##     top200_box - best_actress_win, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.389  -7.552  -0.046   7.292  30.013
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.028e+01  2.030e+02   0.100 0.920467
## title_typeFeature Film         -1.044e+01  4.040e+00  -2.584 0.009986 **
## title_typeTV Movie             -4.900e+00  6.397e+00  -0.766 0.443943
## genreAnimation                  1.885e+00  3.904e+00   0.483 0.629342
## genreArt House & International -6.374e+00  3.394e+00  -1.878 0.060838 .
## genreComedy                     6.109e-01  1.814e+00   0.337 0.736474
## genreDocumentary               -5.138e+00  4.331e+00  -1.186 0.235984
## genreDrama                      1.909e+00  1.571e+00   1.215 0.224670
## genreHorror                     1.962e+00  2.672e+00   0.734 0.462952
## genreMusical & Performing Arts  7.491e-01  3.739e+00   0.200 0.841281
## genreMystery & Suspense         7.395e-01  2.000e+00   0.370 0.711750
## genreOther                      1.177e+00  3.105e+00   0.379 0.704663
## genreScience Fiction & Fantasy -9.454e-01  4.108e+00  -0.230 0.818057
## thtr_rel_year                  -1.887e-01  5.576e-02  -3.384 0.000758 ***
## dvd_rel_year                    1.930e-01  1.268e-01   1.523 0.128360
## imdb_rating                     9.152e+00  5.965e-01  15.343  < 2e-16 ***
## imdb_num_votes                 -1.358e-05  4.704e-06  -2.887 0.004022 **
## critics_ratingFresh            -8.204e+00  1.347e+00  -6.090 1.98e-09 ***
## critics_ratingRotten           -4.156e+01  1.455e+00 -28.570  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 624 degrees of freedom
## Multiple R-squared:  0.8566, Adjusted R-squared:  0.8525
## F-statistic: 207.1 on 18 and 624 DF,  p-value: < 2.2e-16
```

```
vif(data.frame(mod_data2[,c("thtr_rel_year","dvd_rel_year","imdb_rating","imd
b_num_votes")]))
```

```
##        Variables      VIF
## 1  thtr_rel_year 1.812658
## 2   dvd_rel_year 1.774480
## 3    imdb_rating 1.136171
## 4 imdb_num_votes 1.160625
```

Now removing “dvd_rel_year” variable..

```
# removing dvd_rel_yearvariable..

m15 <- lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom-best_di
r_win-top200_box-best_actress_win-dvd_rel_year, data = mod_data)

summary(m15)

##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month - best_pic_nom - best_dir_win -
##     top200_box - best_actress_win - dvd_rel_year, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.737  -7.783   0.017   7.499  29.694
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.997e+02  8.684e+01   3.451 0.000595 ***
## title_typeFeature Film         -1.073e+01  4.040e+00  -2.656 0.008116 **
## title_typeTV Movie             -5.514e+00  6.390e+00  -0.863 0.388519
## genreAnimation                  2.048e+00  3.907e+00   0.524 0.600355
## genreArt House & International -5.955e+00  3.386e+00  -1.759 0.079136 .
## genreComedy                     5.397e-01  1.816e+00   0.297 0.766360
## genreDocumentary               -5.045e+00  4.336e+00  -1.164 0.245011
## genreDrama                      1.954e+00  1.572e+00   1.242 0.214576
## genreHorror                     2.273e+00  2.667e+00   0.852 0.394334
## genreMusical & Performing Arts  8.816e-01  3.742e+00   0.236 0.813822
## genreMystery & Suspense         7.713e-01  2.002e+00   0.385 0.700236
## genreOther                      1.415e+00  3.104e+00   0.456 0.648763
## genreScience Fiction & Fantasy -8.221e-01  4.111e+00  -0.200 0.841569
## thtr_rel_year                  -1.345e-01  4.296e-02  -3.131 0.001824 **
## imdb_rating                     9.061e+00  5.941e-01  15.251  < 2e-16 ***
## imdb_num_votes                 -1.324e-05  4.704e-06  -2.814 0.005048 **
## critics_ratingFresh            -8.083e+00  1.346e+00  -6.004 3.27e-09 ***
## critics_ratingRotten           -4.169e+01  1.454e+00 -28.675  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 625 degrees of freedom
```

```
## Multiple R-squared:  0.8561, Adjusted R-squared:  0.8522
## F-statistic: 218.7 on 17 and 625 DF,  p-value: < 2.2e-16
```

Removing "genre"..

```
m16 <-  lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
    audience_score - runtime -
mpaa_rating-audience_rating-dvd_rel_month-thtr_rel_month-best_pic_nom-best_di
r_win-top200_box-best_actress_win-dvd_rel_year-genre, data = mod_data)

summary(m16)
```

```
##
## Call:
## lm(formula = critics_score ~ . - best_pic_win - best_actor_win -
##     audience_score - runtime - mpaa_rating - audience_rating -
##     dvd_rel_month - thtr_rel_month - best_pic_nom - best_dir_win -
##     top200_box - best_actress_win - dvd_rel_year - genre, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.714  -7.614   0.008   7.998  29.240
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.055e+02  8.547e+01   3.574 0.000378 ***
## title_typeFeature Film -5.105e+00  1.780e+00  -2.867 0.004278 **
## title_typeTV Movie       8.250e-01  5.171e+00   0.160 0.873290
## thtr_rel_year           -1.396e-01  4.224e-02  -3.306 0.001001 **
## imdb_rating              9.081e+00  5.606e-01  16.198  < 2e-16 ***
## imdb_num_votes          -1.302e-05  4.554e-06  -2.859 0.004395 **
## critics_ratingFresh     -8.151e+00  1.336e+00  -6.101 1.83e-09 ***
## critics_ratingRotten    -4.195e+01  1.441e+00 -29.119  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 635 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8519
## F-statistic: 528.5 on 7 and 635 DF,  p-value: < 2.2e-16
```

We see that our Adjusted R squared values decreses by small value that is 0.0003 so we will remove genre from our model as it is not adding significant to our regression model..

```
# Our final model..

final <-
lm(critics_score~title_type+thtr_rel_year+imdb_rating+imdb_num_votes+critics_
rating,
```

```
            data = mod_data)
summary(final)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + thtr_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating, data = mod_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.714  -7.614   0.008   7.998  29.240
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.055e+02  8.547e+01   3.574 0.000378 ***
## title_typeFeature Film -5.105e+00  1.780e+00  -2.867 0.004278 **
## title_typeTV Movie       8.250e-01  5.171e+00   0.160 0.873290
## thtr_rel_year           -1.396e-01  4.224e-02  -3.306 0.001001 **
## imdb_rating              9.081e+00  5.606e-01  16.198  < 2e-16 ***
## imdb_num_votes          -1.302e-05  4.554e-06  -2.859 0.004395 **
## critics_ratingFresh     -8.151e+00  1.336e+00  -6.101 1.83e-09 ***
## critics_ratingRotten    -4.195e+01  1.441e+00 -29.119  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 635 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8519
## F-statistic: 528.5 on 7 and 635 DF,  p-value: < 2.2e-16
```

Part 5 - Cross Validation

We will see how our existing model works on new data. What we will do is divide the data into training and test data, we will build model around training data and see the performance on test set..

```
# Dividing the data into training and test set

 set.seed(123)
split = sample.split(mod_data$critics_score, SplitRatio = 0.70)
training_set = subset(mod_data,split == T)
test_set = subset(mod_data, split == F)
```

Now I decided to give 70% of the data to training and remaining 30% to the test set.

```
final_train <-
lm(critics_score~title_type+thtr_rel_year+imdb_rating+imdb_num_votes+critics_
rating,
```

```
            data = training_set)
summary(final_train)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + thtr_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7765  -7.7237  -0.0077   7.6511  29.2127
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.115e+02  9.941e+01   4.139 4.18e-05 ***
## title_typeFeature Film -5.033e+00  2.114e+00  -2.381 0.017674 *
## title_typeTV Movie       7.541e+00  5.834e+00   1.293 0.196825
## thtr_rel_year           -1.924e-01  4.921e-02  -3.910 0.000107 ***
## imdb_rating              9.053e+00  6.568e-01  13.784  < 2e-16 ***
## imdb_num_votes          -1.302e-05  5.519e-06  -2.360 0.018718 *
## critics_ratingFresh     -8.971e+00  1.585e+00  -5.661 2.72e-08 ***
## critics_ratingRotten    -4.216e+01  1.714e+00 -24.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.88 on 442 degrees of freedom
## Multiple R-squared:  0.8559, Adjusted R-squared:  0.8536
## F-statistic:   375 on 7 and 442 DF,  p-value: < 2.2e-16
```

The summary is good, now let's check this on our test data..

```
fitted <- predict(final_train, test_set)
```

Now let's check the efficiency of the model on test set...

```
deviation <- test_set$critics_score-fitted ## deviation

per_deviation <- (test_set$critics_score-fitted)/test_set$critics_score  ##%
deviation

abs_deviation <- abs(test_set$critics_score-fitted)/test_set$critics_score
##absolute deviation

mean(abs(test_set$critics_score-fitted)/test_set$critics_score) ##Mean abs %
error
```

```
## [1] 0.3466112
```

```
#MAPE - mean absolute percentage deviation
```

The value of MAPE is 0.34 which is not big, this can be reduced by using other methods, so we can say that our model is good. But we cannot stop here. For making a regression model we always make make some assumptions. We need to check that the assumptions are fulfilled.

Part 6 - Interpretations of model coefficients:

Consider the intercept in the final model. This shows that if a documentary movie which does not show it's release year, it's Imdb rating, imdb number of votes and it is being rated by Certified fresh, the critics rating will be 411 which is vague and not possible.

For categorical variables like "title_type" and "critics_rating", one level of the variable is kept 0 which means that the level which is made 0. And the interpretation of other levels is made in reference to the factor made 0.

Considering the theatre release year variables, we can interpret that keeping rest of the variables constant, for a unit increase in year the critic score of rotten tomatoes decreases by about-1.924.

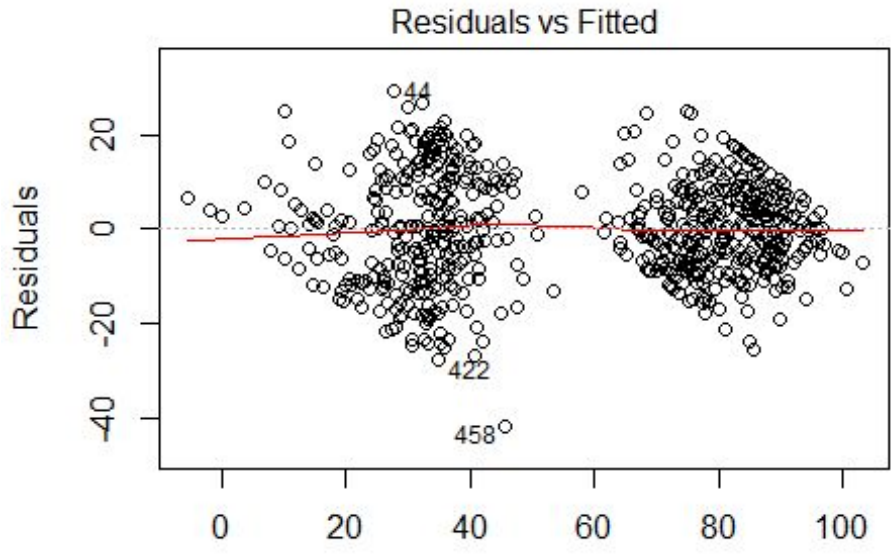
Consider the imdb_num_votes variables, keeping rest of the variables constant, for unit increase in IMDB voters the critic rating is likely to decrease by 1.302e-05, which is vey small.

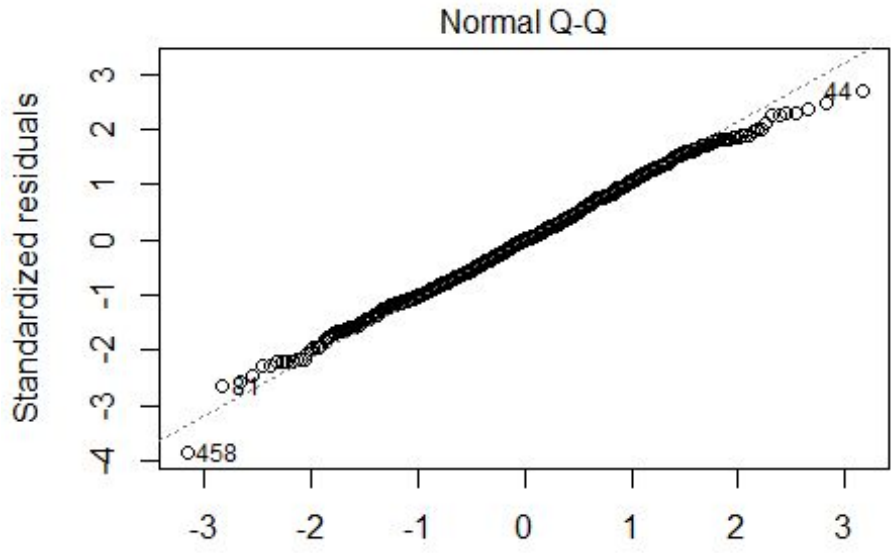
Part 7 - Model Diagnostics:

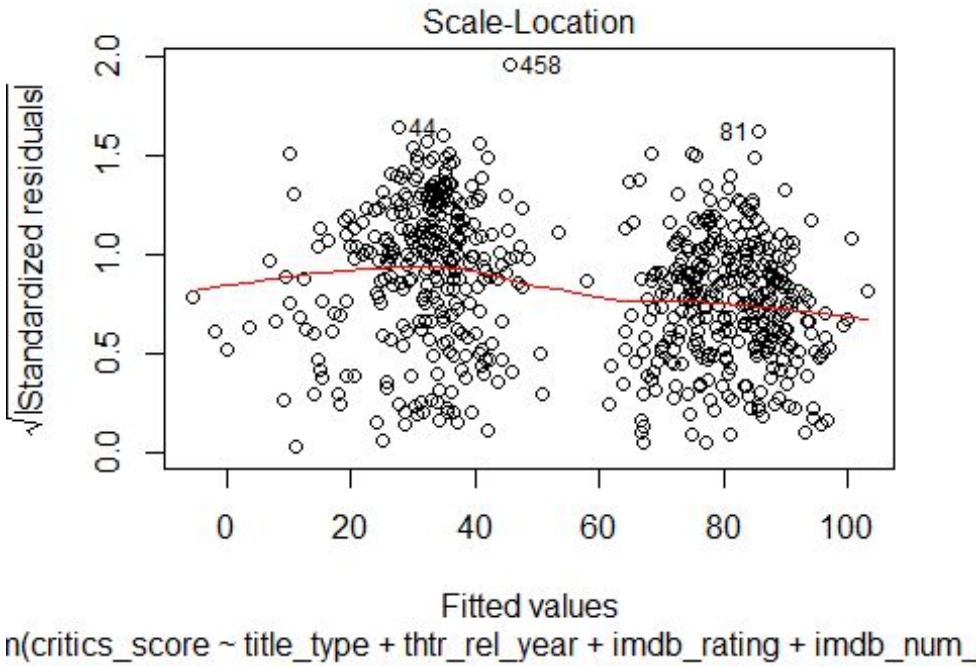
So we made our model. Now we check conditions required for multiple regression to be mapped valid.

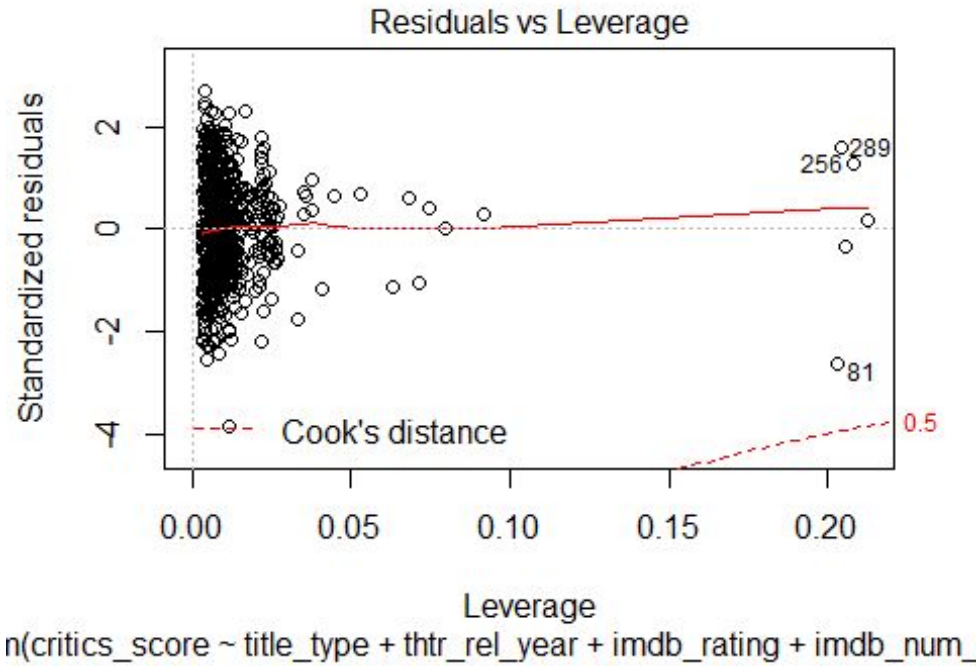
1. The first condition is the linear relationship between numerical x and response variable. We can check this using residual plot with x variable(numerical).

```
plot(final)
```


Residuals vs Fitted
Residuals
Fitted values
n(critics_score ~ title_type + thtr_rel_year + imdb_rating + imdb_num_
44
422
458
20
0
-20
-40
0
20
40
60
80
100

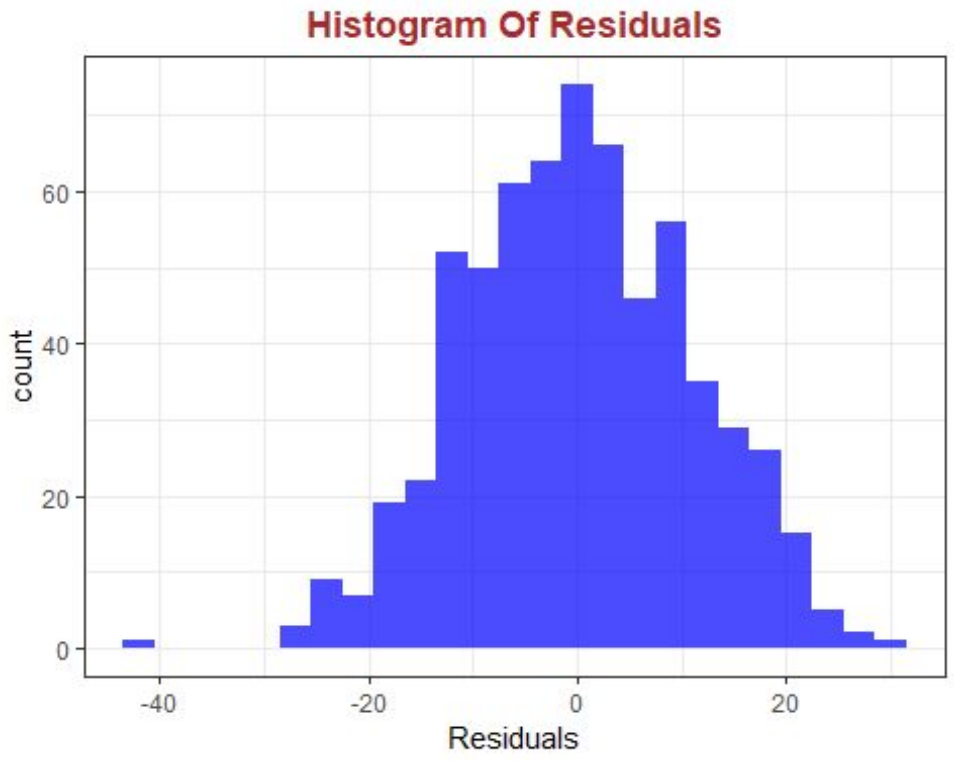

Normal Q-Q
Standardized residuals
Theoretical Quantiles
n(critics_score ~ title_type + thtr_rel_year + imdb_rating + imdb_num_
44
458
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3





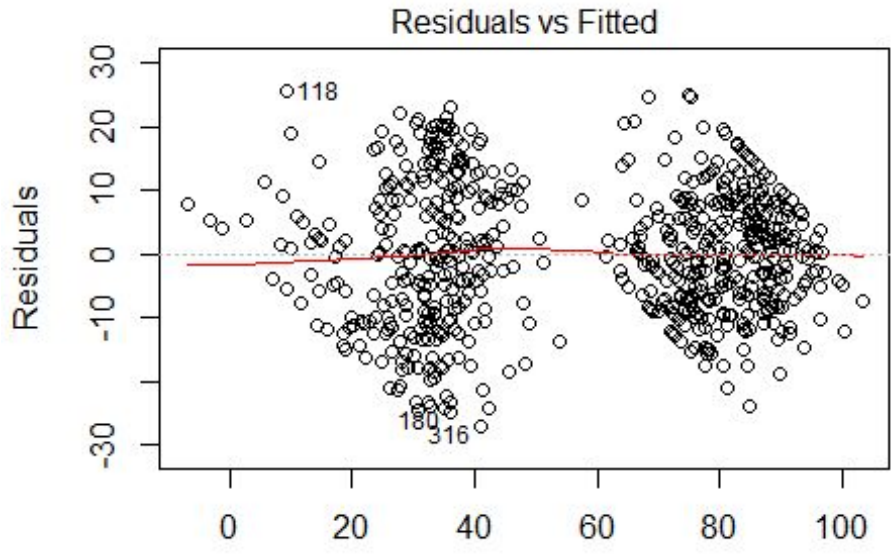
From Residuals vs fitted values we can see that the data point 44 is very influential, we can see the effect more clearly by plotting the histogram..

```
res <- residuals(final)
res <- as.data.frame(res)
ggplot(res,aes(res)) +  geom_histogram(fill='blue',alpha=0.7, binwidth = 3) +
xlab("Residuals") + ggtitle("Histogram Of Residuals") + theme_bw() +
theme(plot.title = element_text(hjust = 0.5, colour = "Brown", face =
"bold"))
```

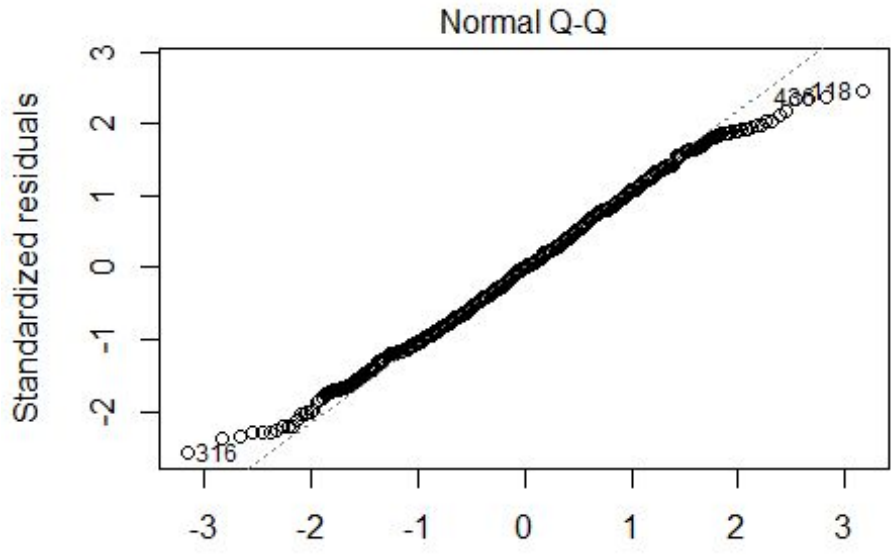


Our residual plot becomes left skewed by some points as shown in the plot. We will remove the data point at 44 and 458 as they are not influential also..

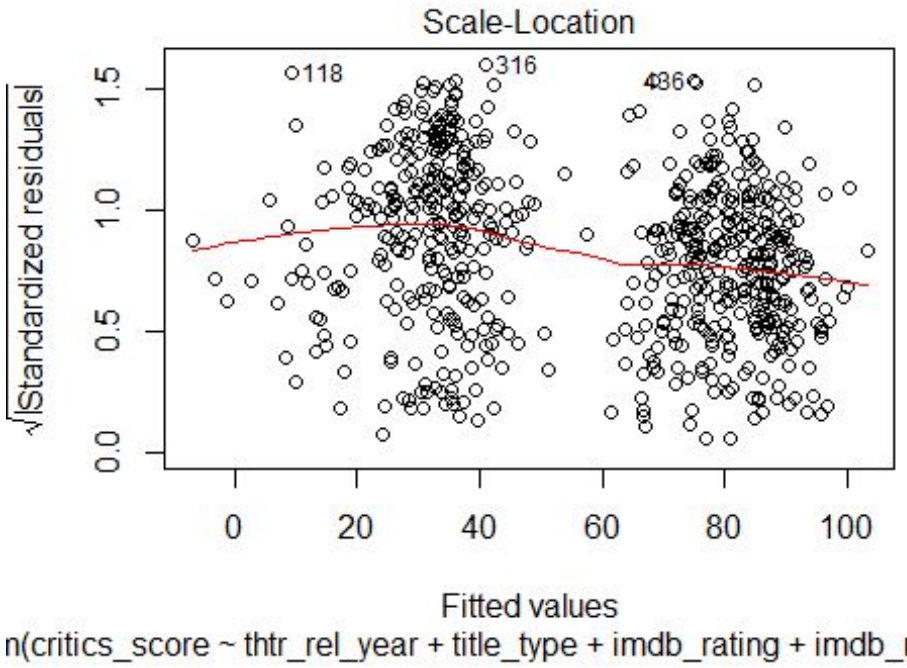
```
mod_data5 <- mod_data[c(-44,-458),]
# I have removed some data points by checking the plots previously
mod_data6 <- mod_data5[c(-80,-77),]


final1 <- lm(critics_score~thtr_rel_year+
title_type+imdb_rating+imdb_num_votes+critics_rating,
             data = mod_data6[c(-57,-419),])

plot(final1)
```

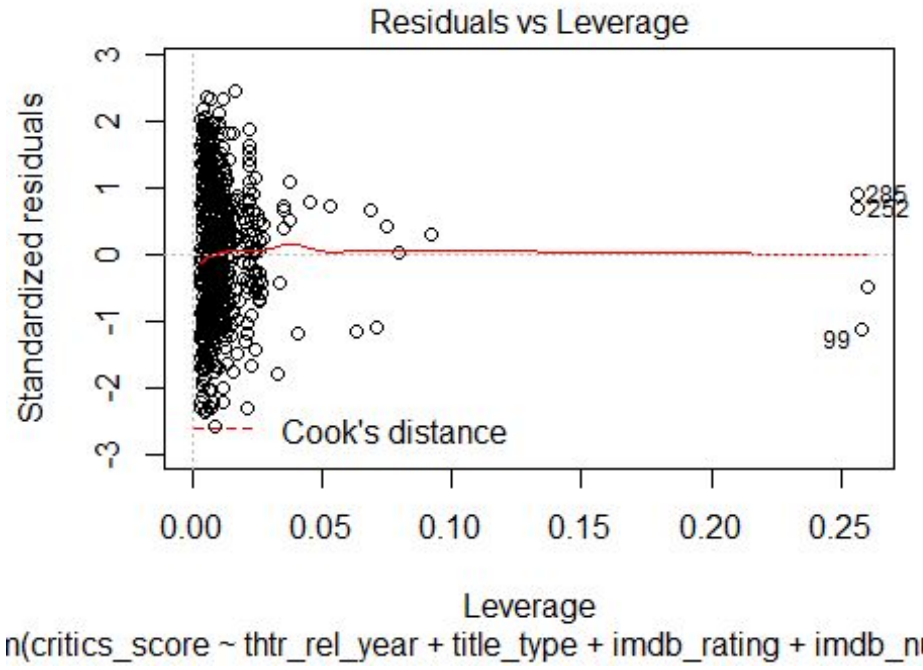


n(critics_score ~ thtr_rel_year + title_type + imdb_rating + imdb_num_



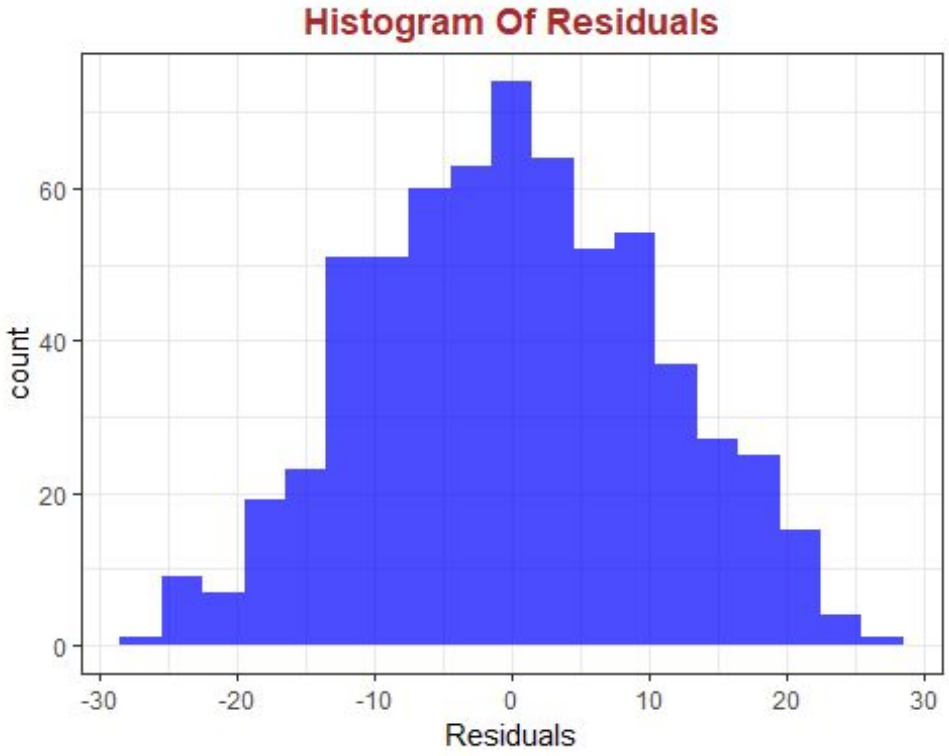
n(critics_score ~ thtr_rel_year + title_type + imdb_rating + imdb_num_






```
res <- residuals(final1)
res <- as.data.frame(res)
ggplot(res,aes(res)) +  geom_histogram(fill='blue',alpha=0.7, binwidth = 3) +
```

```
xlab("Residuals") + ggtitle("Histogram Of Residuals") + theme_bw() +
theme(plot.title = element_text(hjust = 0.5, colour = "Brown", face =
"bold"))
```



1. Now we by seeing different plots we can see that residuals and fitted values plot has values has a random scatter except between 45 to 55 as my model do not predict the values between the given interval, so we can say that our model is somewhat homoscedasticity..

2. Looking at the Normal Q-Q So we can see there is random scatter around 0, and by plotting histogram of residuals we see that the distribution is normally distributed centered around mean..

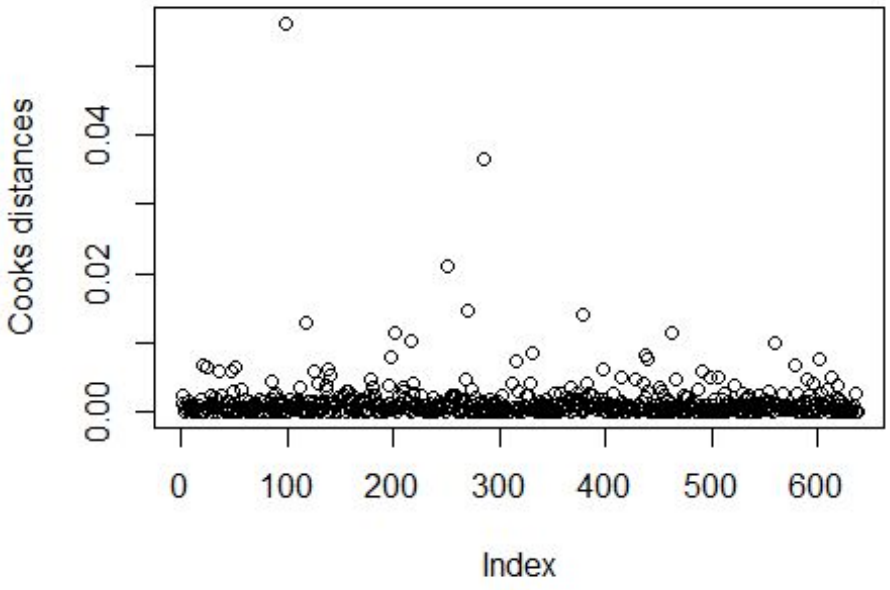
The plot between fitted and standardised residual is same except, here the residuals are standardized, and the plot is random which is good for our model..

3. The residuals vs levarage shows the data points from which our model affects significantly..

There is one criteria called cooks distance from which we can find the data point which is very influential..

The data point having cooks distance greater than 0.8 should be removed from our model..

```r
# plotting cooks distance
cook = cooks.distance(final1)
plot(cook,ylab="Cooks distances")
```



We can see that the cooks distance for all data points is very low, so we dont need to remove any data point from our model.

4. Now we check for autocorrelation in the observations.. There are many ways to check it, for example making Auto correlation plot, by durbin watson test or by using runs.test.

I am using durbin watson test for making inference..

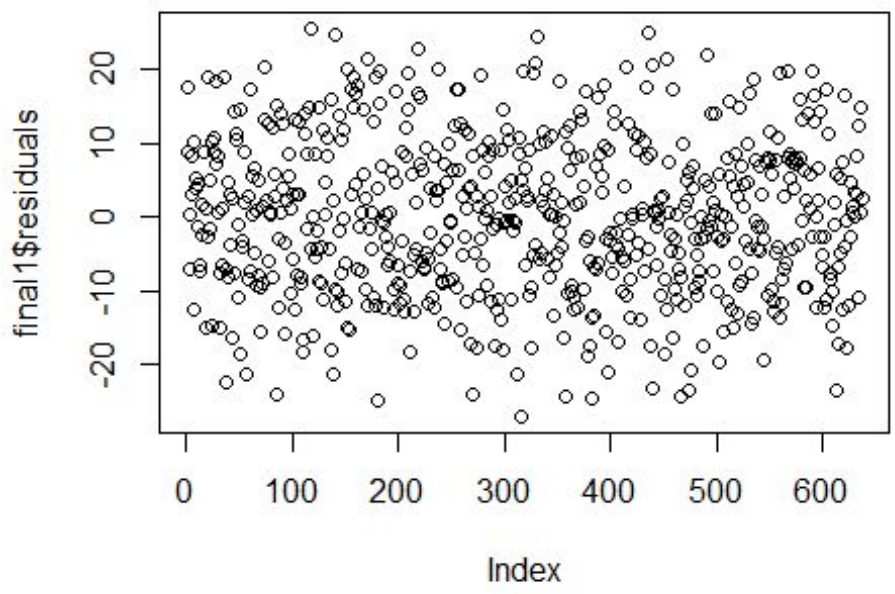
```r
# we will need lmtest package for it...

dwtest(final1)
```

```
##
##  Durbin-Watson test
##
## data:  final1
## DW = 2.0063, p-value = 0.532
## alternative hypothesis: true autocorrelation is greater than 0
```

The hypothesis are : Null - True autocorrelation is 0. Alternative - True autocorrelation is greater than 0.

Since the p-value is very high than 0.05 so we fail to reject the null hypothesis..

```
plot(final1$residuals)
```



From this we check that the residuals are randomly scattered so we can say that there is no autocorrelation.

5. Multicollinearity - There may be a possibility that the predictor variables are themselves correlated among themselves, this will be a problem as our model will be added with unnecessary redundancies. This can be handled by checking vif. We have already checked vif for our variables, I am showing it once again for the sake of completing this assumptions section. The vif greater than 4 is undesirable..

```
vif(data.frame(mod_data6[c(-57,-419),c("thtr_rel_year","title_type",
"imdb_num_votes","imdb_rating","critics_rating")]))
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors

## Warning in Ops.factor(r, 2): '^' not meaningful for factors

## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored

## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

```
## Warning in Ops.factor(r, 2): '^' not meaningful for factors
```

```
##        Variables      VIF
## 1  thtr_rel_year 1.144649
## 2     title_type       NA
## 3 imdb_num_votes 1.421460
## 4    imdb_rating 1.969039
## 5 critics_rating       NA
```

As title_type and critics_rating are categorical variables, we compute vif only for numeric values, so there is NA in the output. As we can see the vif is less than 4 for all the numeric predictors, so we can say that there is no problem of multicollinearity.

Part 7 - Let's check our model on any movie

Let's see the maximum value predicted by model.

```
# Maximum value of prediction
max(final1$fitted.values)
```

```
## [1] 103.2646
```

But when we look at the maximum value of our prediction, we see that it is 103.2646, we know that rotten tomatoes maximum score is 100. So we need to add certain restriction to our model so that our prediction doesn't exceeds 100. We can do it by making simple fucntion.

```
predict_values <- data.frame(final1$fitted.values)
colnames(predict_values) <- c("Predict")
greater_hundred <- function(x){
    if  (x > 100){
        return(100)
    }else{
        return(x)
    }
}

predict_values$Predict <- sapply(predict_values$Predict, greater_hundred)
max(predict_values$Predict)
```

```
## [1] 100
```

Now I'm going to predict the rotten tomatoes rating of one of my favourite movie of 2016 "Deadpool".

and regarding imdb_num_votes can be found at
http://www.imdb.com/title/tt1431045/ratings?ref_=tt_ov_rt

dvd is released on May 2016,can be found by clicking this link
https://www.imdb.com/title/tt1431045/?ref_=tt_rt

Deadpool is certified fresh which can be found on
https://www.rottentomatoes.com/m/deadpool/

Now we have information regarding all the variables needed to make our model, we will put all in a data frame,

```
# Taking information from a particualar movie which is not in the data set..

newmovie <- data.frame(thtr_rel_year = 2016, title_type = "Feature Film",
imdb_rating = 8, imdb_num_votes = 747563, critics_rating = "Certified Fresh")
```

Now let's see what our model predicts.

```
# predicting the value of critics score for the newmovie..

predict(final1, newmovie)
```

```
##        1
## 81.66647
```

So our predicted value for rotten tomatoes critics rating is 81.66647 while the actual critic score which is available on rotten romatoes site is 83 as given on the site. So we can say that our model can approximately predicts the critic score of rotten tomatoes.

We can also construct a prediction interval around this prediction, which will provide a measure of uncertainty around the prediction.

```
#predicting newmovie confidence interval for the value of critics score..

predict(final1, newmovie, interval = "confidence", level = 0.95)
```

```
##        fit     lwr      upr
## 1 81.66647 76.4622 86.87074
```

The above statement says that "We are 95% confident that the movie"DEADPOOL" will get critics score on average between 76.4622 and 86.87074 by Rotten tomatoes."

Part 6: Conclusion

Thus we have found the factors which decides the success of the movie,like Imdb Rating, title type etc, but we need more parameters(variables) to make more accurate prediction of the review as for now we can only approximate our findings based on the variables given in the data set. Like say "Box Office" and "Budget" can also play a very important role in predicting the crtics score of movie, like wise there are many.

The model can be used to predict the success rate of movie and by adding some input variables in the data set we can improve the performance of the model.