Modeling and prediction for movies

Made By – Varshit Dubey (CoE Pune)

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(ggthemes)
library(corrgram)
library(corrplot)
library(caTools)
```

Load data

```
load("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/" ...
```

```
# 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
```

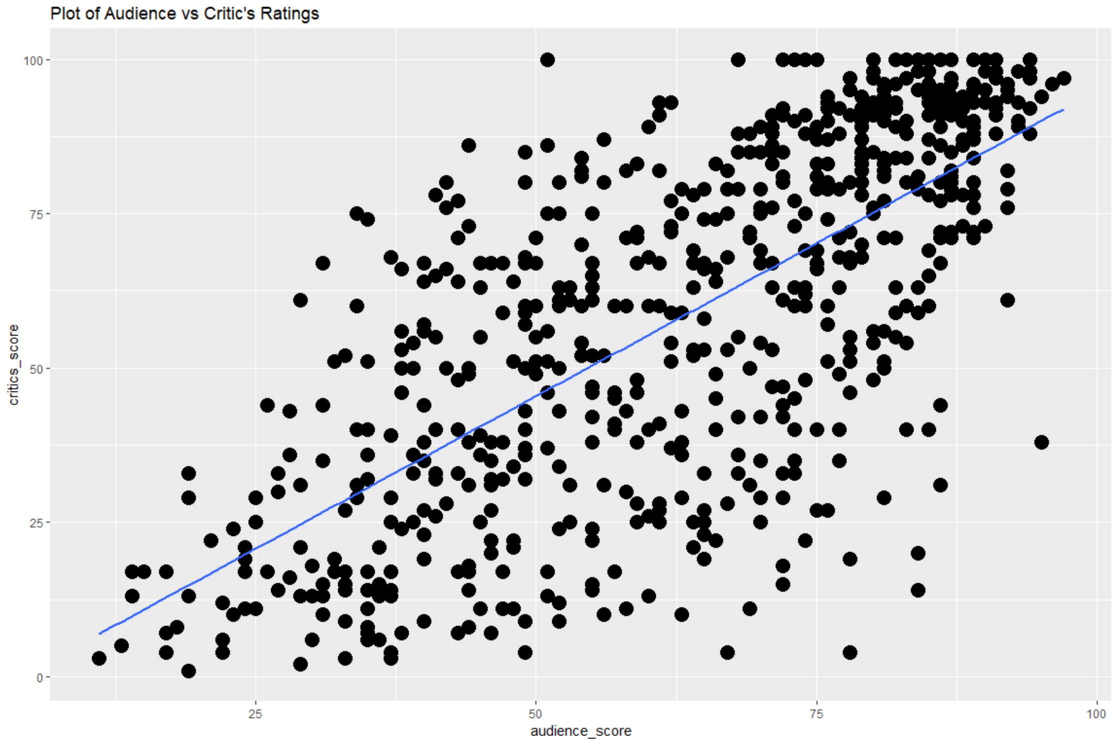
This table shows the average rating given by critics and audience based on genre. As you can see from the table there is difference in rating. So we can say that audience and critics and audience have different taste of movies.

```
# 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
```

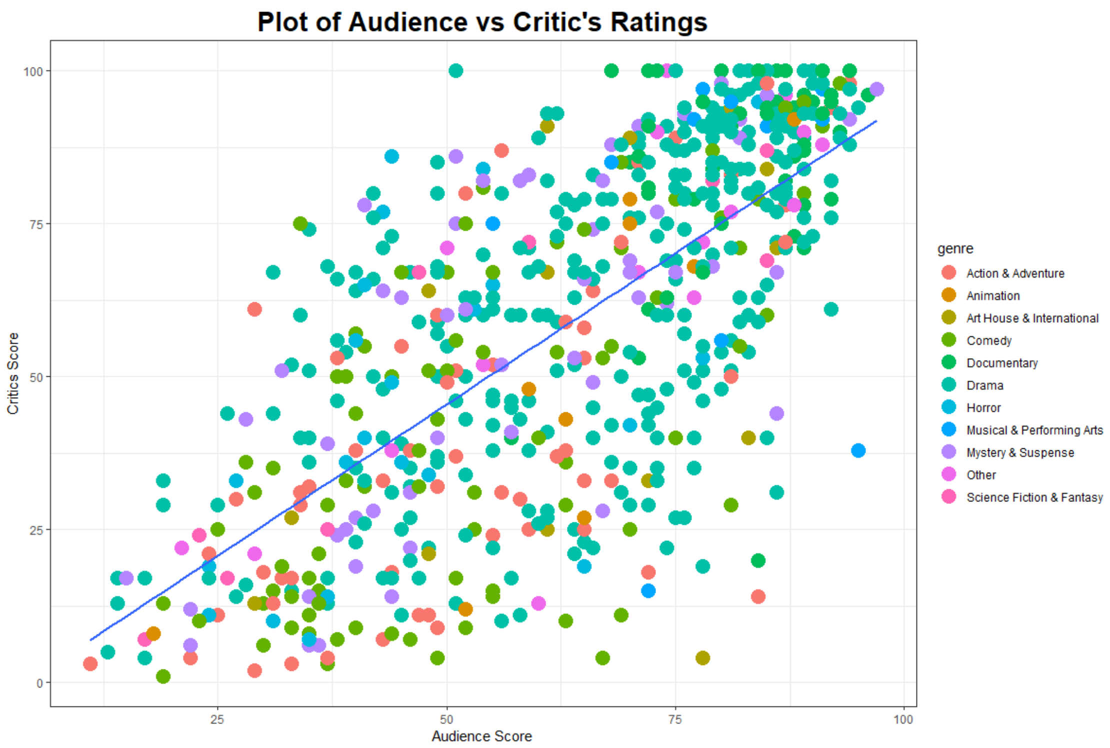
And the second table shows the differece in audience and critics rating on rotten tomatoes.The above result shows that Audience tend to like “Action and adventure” movies while Critics give them low ratings on average.

```
# 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")
```

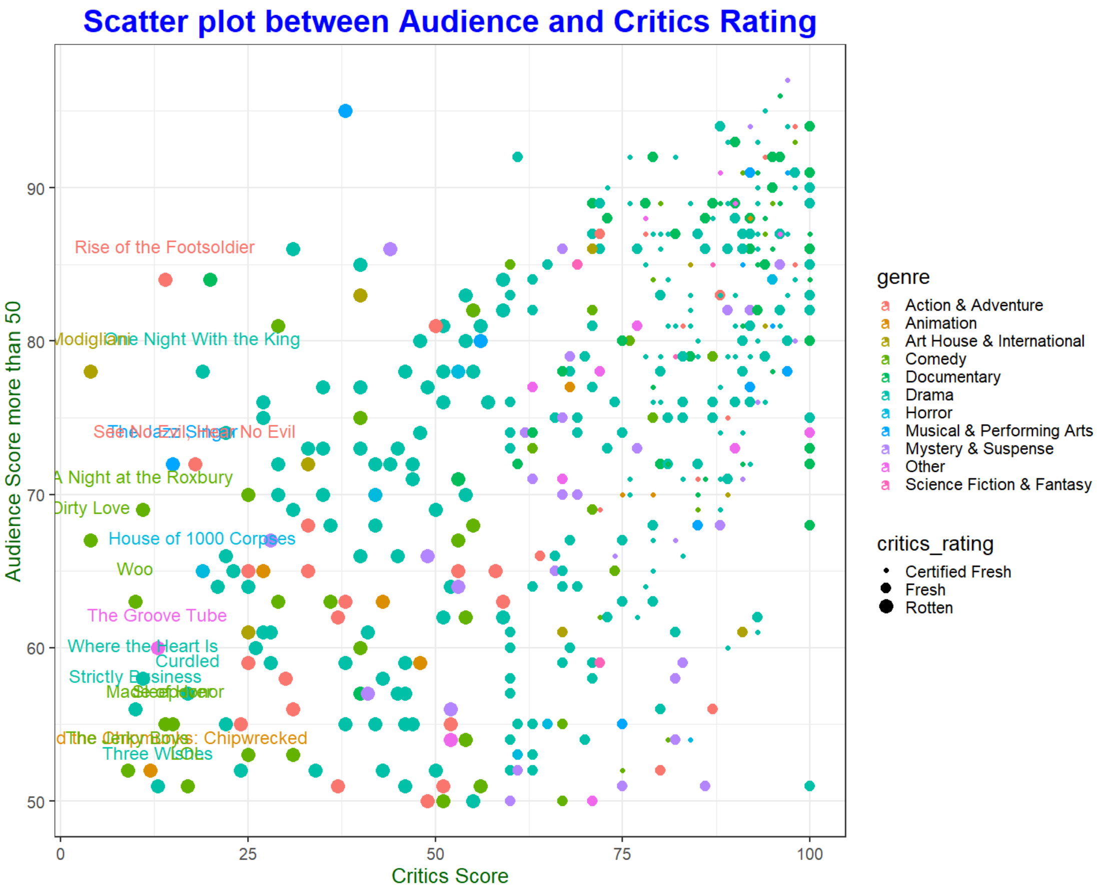


The above plot shows there is positive linear relation between audience and critic scores. Let's make this plot more attractive.

```
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.

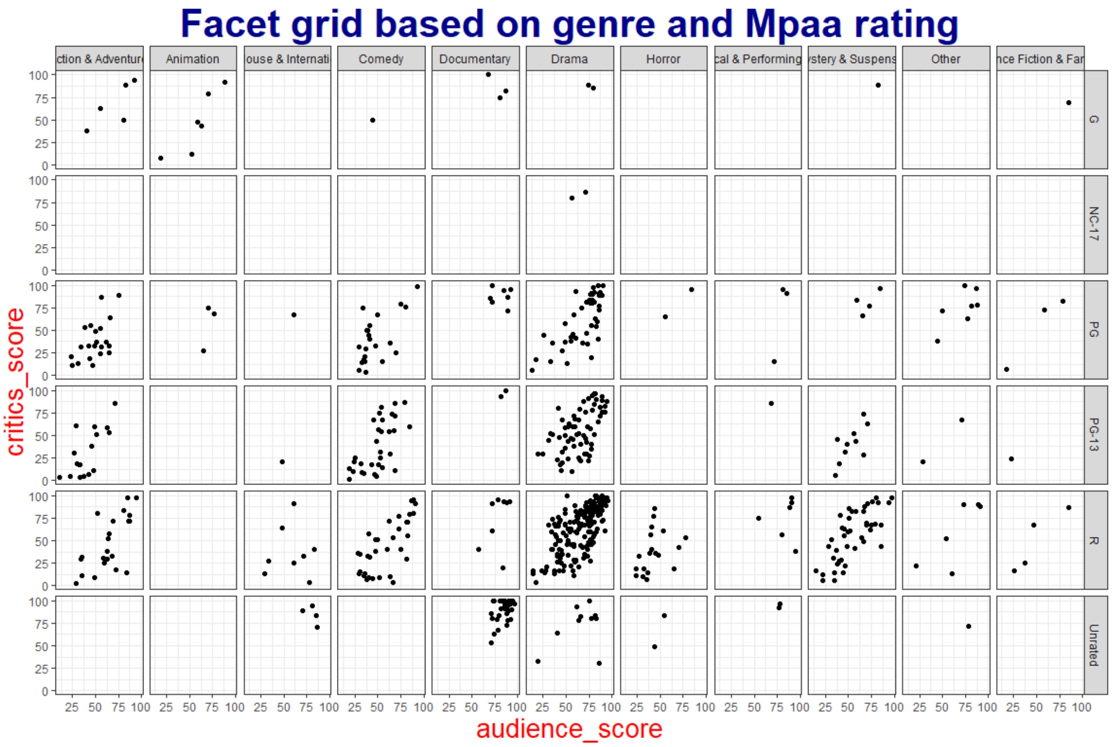
And I plot the scatterplot, between critics_score an audience_score, and also added size and colour to make it more beautiful. Ignore the warning.

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”.

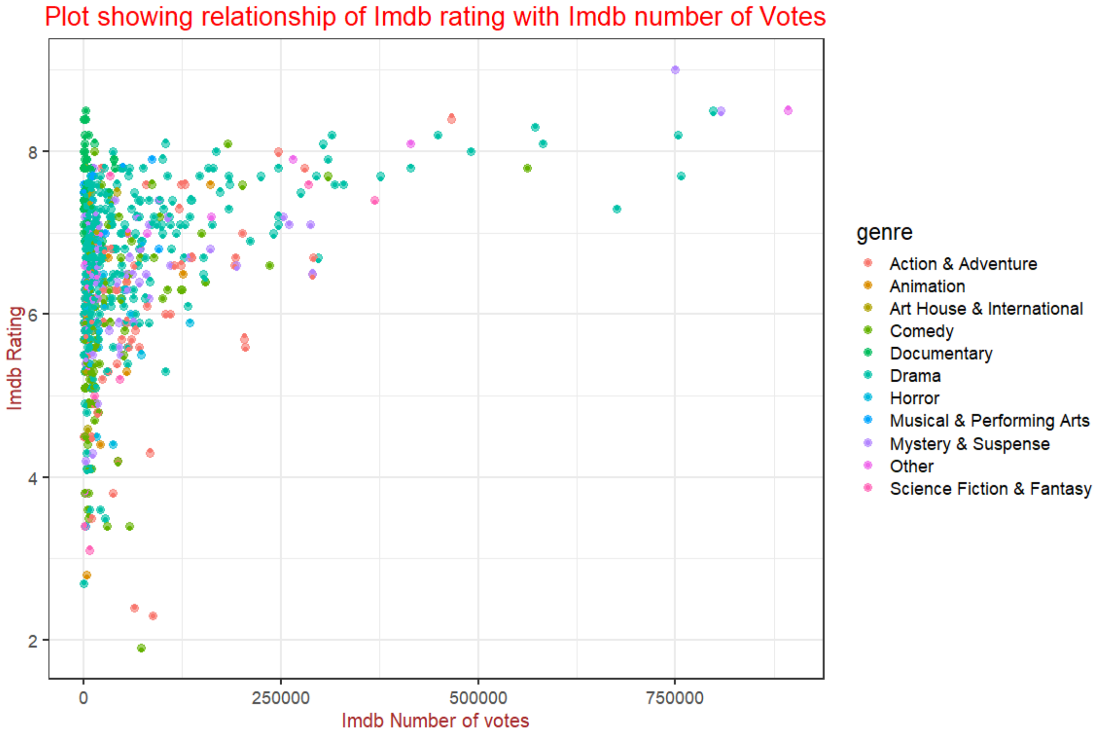
Now we try to make a scatterplot which shows the plot based on individual levels of categorical variables, like a grid.

```
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))
```



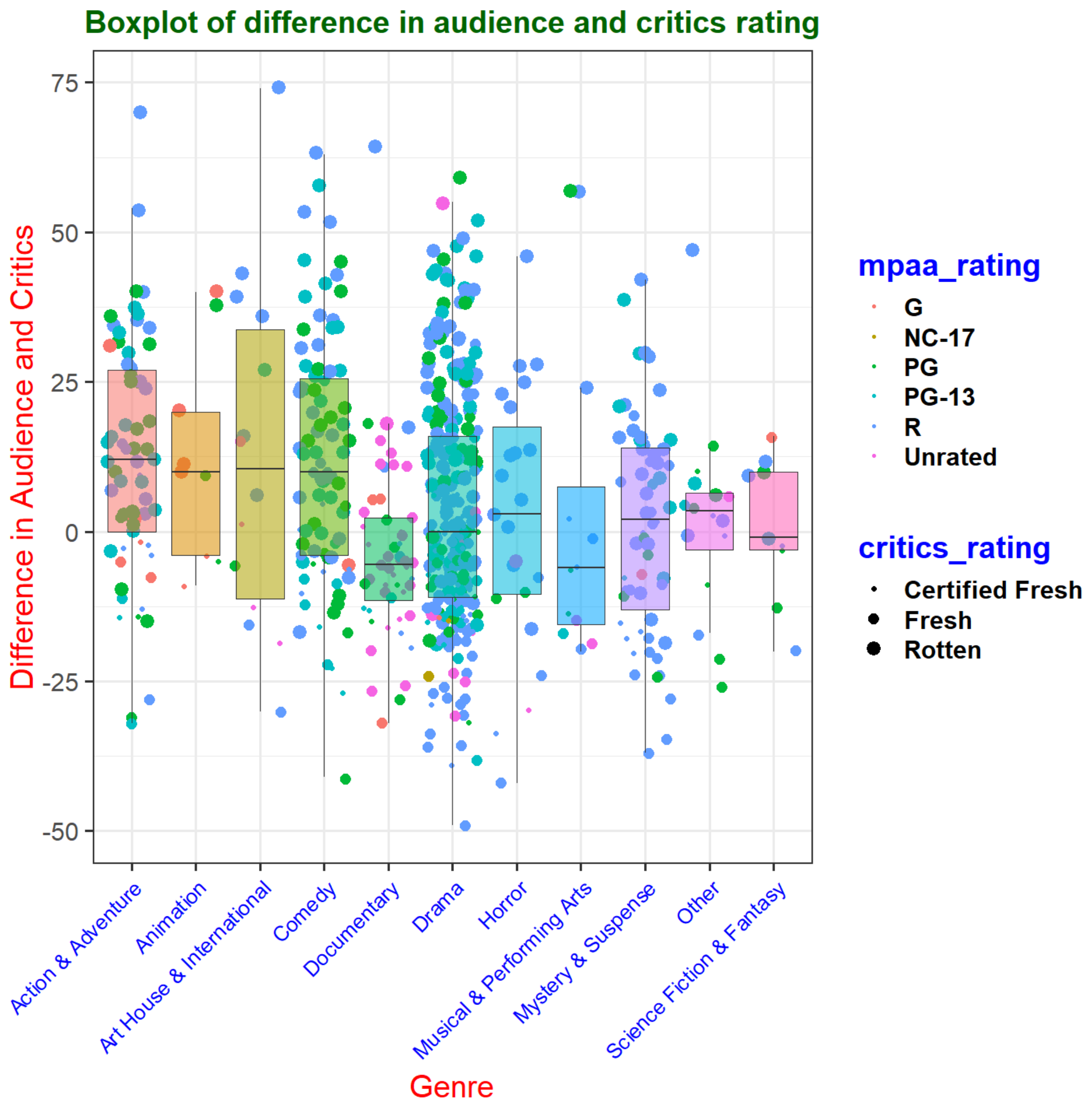
Now let’s try to make scatter plot with another variables..

```
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.
```



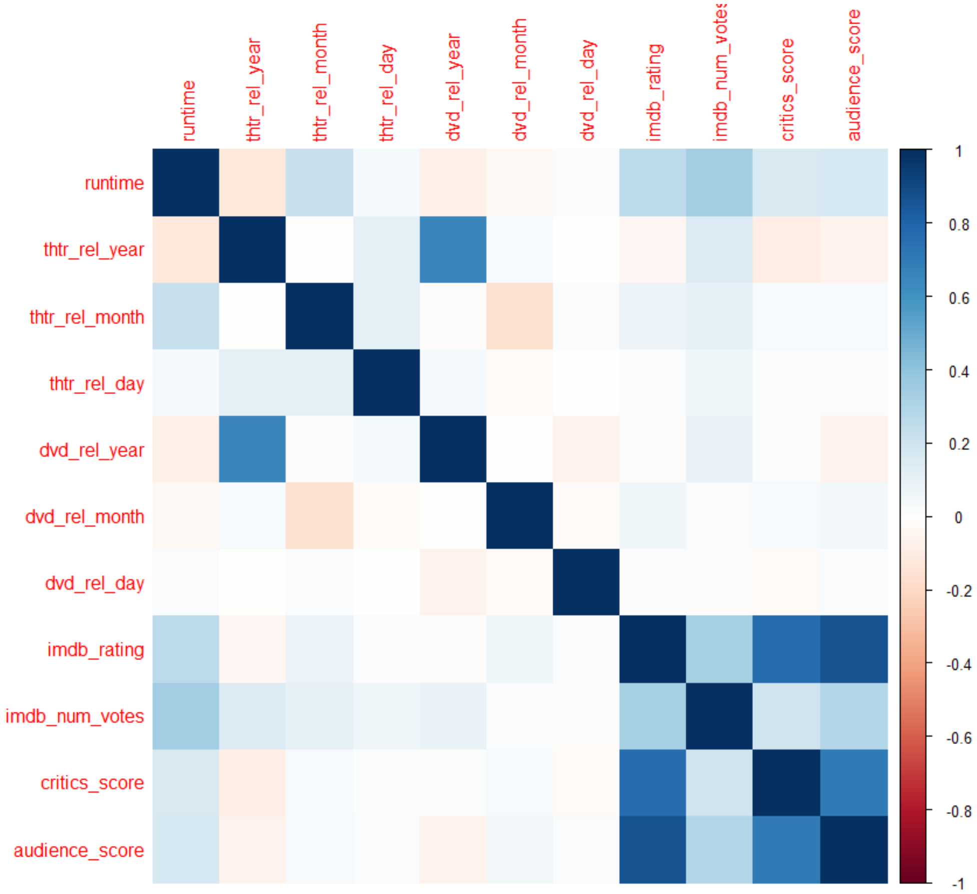
The above boxplot shows the distribution of observations of difference in audience and critics score, based on different genres. We can see that audience tend to score the movie more than critics for most of the genre as their the median of many boxplots is more than 0, but for some genres like “Documentary”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

# 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
```

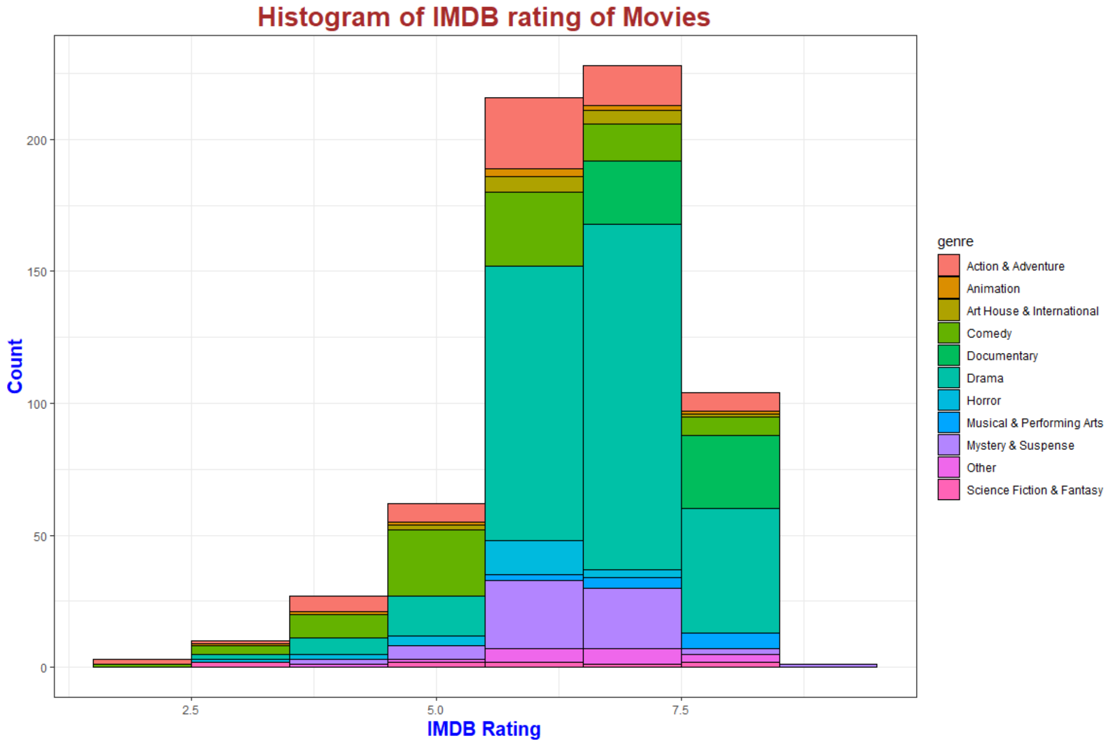
```
# Making correlation plot.

corrplot(cor.data,method='color')
```



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.

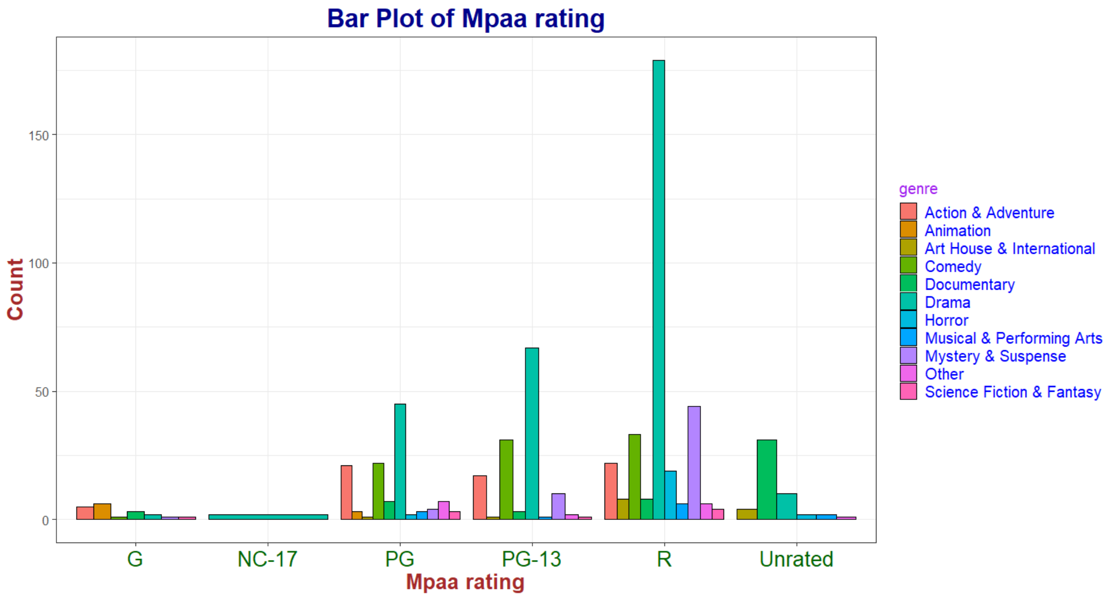
```
ggplot(data = movies, aes(x = imdb_rating)) +
  geom_histogram(mapping = aes(fill = genre), colour = "black", binwidth = 1)
+
  theme_bw() +
  labs(title = "Histogram of IMDB rating of Movies", x = "IMDB Rating", y =
"Count") +
  theme(plot.title = element_text(size = 20, face = "bold", color = "Brown",
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.
mod_data <- movies %>% select(-c(director:rt_url), -title, -thtr_rel_day, -
dvd_rel_day, -studio)

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

```r
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)
```

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.

```r
#assumed model

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

summary(m1)

## 
## Call:
## lm(formula = critics_score ~ ., data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.6718  -7.5658  -0.0195   6.9793  28.4505 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                     2.191e+02  2.486e+02   0.881  0.37865  
## title_typeFeature Film         -1.163e+01  4.719e+00  -2.463  0.01417 *
## title_typeTV Movie             -1.761e+00  7.442e+00  -0.237  0.81303  
## genreAnimation                 -2.683e-01  5.850e+00  -0.046  0.96344  
## genreArt House & International -4.716e+00  4.174e+00  -1.130  0.25919  
## genreComedy                     1.242e+00  2.191e+00   0.567  0.57123  
## genreDocumentary               -7.752e+00  5.027e+00  -1.542  0.12383  
## genreDrama                      2.457e+00  1.971e+00   1.247  0.21326  
## genreHorror                     1.772e+00  3.515e+00   0.504  0.61448  
## genreMusical & Performing Arts -5.112e+00  6.310e+00  -0.810  0.41834  
## genreMystery & Suspense        -8.277e-01  2.516e+00  -0.329  0.74237  
## genreOther                      1.280e+00  4.001e+00   0.320  0.74923  
## genreScience Fiction & Fantasy -1.622e+00  4.553e+00  -0.356  0.72188  
## runtime                        -2.580e-02  3.745e-02  -0.689  0.49128  
## mpaa_ratingNC-17               -1.647e+00  8.620e+00  -0.191  0.84856  
## mpaa_ratingPG                  -2.064e+00  3.533e+00  -0.584  0.55939  
## mpaa_ratingPG-13               -3.323e+00  3.716e+00  -0.894  0.37166  
## mpaa_ratingR                   -3.338e+00  3.519e+00  -0.949  0.34337  
## mpaa_ratingUnrated              5.825e-01  4.203e+00   0.139  0.88983  
## thtr_rel_year                  -2.047e-01  7.297e-02  -2.805  0.00527 **
## thtr_rel_month                 -1.574e-01  1.570e-01  -1.003  0.31647
## dvd_rel_year                    1.134e-01  1.550e-01   0.731  0.46497
## dvd_rel_month                  -1.582e-01  1.611e-01  -0.982  0.32660
## imdb_rating                     9.326e+00  1.093e+00   8.532 2.73e-16 ***
## imdb_num_votes                 -1.340e-05  6.401e-06  -2.093  0.03697 *
## critics_ratingFresh            -8.592e+00  1.680e+00  -5.116 4.79e-07 ***
## critics_ratingRotten           -4.108e+01  1.849e+00 -22.217  < 2e-16 ***
## audience_ratingUpright         -8.810e-01  2.188e+00  -0.403  0.68742
## audience_score                  5.230e-03  7.723e-02   0.068  0.94604
## best_pic_nomyes                 1.885e+00  3.286e+00   0.574  0.56655
## best_pic_winyes                 4.798e-01  6.152e+00   0.078  0.93787
## best_actor_winyes               6.995e-02  1.635e+00   0.043  0.96590
## best_actress_winyes             7.437e-01  1.797e+00   0.414  0.67922
## best_dir_winyes                 2.204e+00  2.319e+00   0.950  0.34245
## top200_boxyes                   2.818e+00  3.708e+00   0.760  0.44769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.98 on 415 degrees of freedom
## Multiple R-squared:  0.862,  Adjusted R-squared:  0.8507
## F-statistic: 76.23 on 34 and 415 DF,  p-value: < 2.2e-16
```

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

```
# without genre
str(training_set)

## Classes 'tbl_df', 'tbl' and 'data.frame':    450 obs. of  20 variables:
##  $ title_type      : Factor w/ 3 levels "Documentary",..: 2 2 2 2 1 2 2 1
2 2 ...
##  $ genre           : Factor w/ 11 levels "Action & Adventure",..: 6 6 4 6
5 6 6 5 1 4 ...
##  $ runtime         : num  80 101 84 139 78 142 93 88 127 100 ...
##  $ mpaa_rating     : Factor w/ 6 levels "G","NC-17","PG",..: 5 4 5 3 6 4 5
6 3 4 ...
##  $ thtr_rel_year   : num  2013 2001 1996 1993 2009 ...
##  $ thtr_rel_month  : num  4 3 8 10 1 1 11 9 6 9 ...
##  $ dvd_rel_year    : num  2013 2001 2001 2001 2010 ...
##  $ dvd_rel_month   : num  7 8 8 11 4 2 3 1 5 2 ...
##  $ imdb_rating     : num  5.5 7.3 7.6 7.2 7.8 7.2 5.5 7.5 6.8 5.9 ...
##  $ imdb_num_votes  : int  899 12285 22381 35096 333 5016 2272 880 71979
25808 ...
##  $ critics_rating  : Factor w/ 3 levels "Certified Fresh",..: 3 1 1 1 2 3
3 2 1 3 ...
##  $ critics_score   : num  45 96 91 80 91 57 17 90 89 25 ...
##  $ audience_rating : Factor w/ 2 levels "Spilled","Upright": 2 2 2 2 2 2 1
2 2 1 ...
##  $ audience_score  : num  73 81 91 76 86 76 47 89 75 53 ...
##  $ best_pic_nom    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1
...
##  $ best_pic_win    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1
...
##  $ best_actor_win  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 2 1 2 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 2 1
...
```

```
m2 <- lm(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_dir_win + top200_box, data= training_set)

summary(m2)
```

```
##
## Call:
## lm(formula = 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_dir_win + top200_box, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7489  -7.6282  -0.0379   6.9801  28.3741
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.220e+02  2.483e+02   0.894  0.37183
## title_typeFeature Film         -1.161e+01  4.714e+00  -2.462  0.01420 *
## title_typeTV Movie             -1.613e+00  7.426e+00  -0.217  0.82814
## genreAnimation                 -7.248e-02  5.825e+00  -0.012  0.99008
## genreArt House & International -4.697e+00  4.169e+00  -1.127  0.26056
## genreComedy                     1.340e+00  2.176e+00   0.616  0.53847
## genreDocumentary               -7.690e+00  5.020e+00  -1.532  0.12632
## genreDrama                      2.558e+00  1.954e+00   1.309  0.19113
## genreHorror                     1.805e+00  3.510e+00   0.514  0.60744
## genreMusical & Performing Arts -5.121e+00  6.303e+00  -0.812  0.41703
## genreMystery & Suspense        -7.322e-01  2.503e+00  -0.293  0.77004
## genreOther                      1.227e+00  3.995e+00   0.307  0.75895
## genreScience Fiction & Fantasy -1.587e+00  4.548e+00  -0.349  0.72723
## runtime                        -2.419e-02  3.721e-02  -0.650  0.51599    
## mpaa_ratingNC-17               -1.727e+00  8.609e+00  -0.201  0.84112    
## mpaa_ratingPG                  -2.025e+00  3.529e+00  -0.574  0.56642    
## mpaa_ratingPG-13               -3.322e+00  3.712e+00  -0.895  0.37138    
## mpaa_ratingR                   -3.331e+00  3.515e+00  -0.948  0.34385    
## mpaa_ratingUnrated              6.207e-01  4.197e+00   0.148  0.88252    
## thtr_rel_year                  -2.044e-01  7.289e-02  -2.804  0.00529 ** 
## thtr_rel_month                 -1.585e-01  1.568e-01  -1.011  0.31250    
## dvd_rel_year                    1.116e-01  1.548e-01   0.721  0.47150    
## dvd_rel_month                  -1.605e-01  1.608e-01  -0.998  0.31874    
## imdb_rating                     9.347e+00  1.091e+00   8.568  < 2e-16 ***
## imdb_num_votes                 -1.342e-05  6.395e-06  -2.098  0.03649 *  
## critics_ratingFresh            -8.612e+00  1.677e+00  -5.135 4.35e-07 ***
## critics_ratingRotten           -4.111e+01  1.847e+00 -22.262  < 2e-16 ***
## audience_ratingUpright         -8.640e-01  2.185e+00  -0.395  0.69279    
## audience_score                  3.314e-03  7.702e-02   0.043  0.96570    
## best_pic_nomyes                 2.074e+00  3.250e+00   0.638  0.52381    
## best_pic_winyes                 5.542e-01  6.143e+00   0.090  0.92815    
## best_actor_winyes               8.895e-02  1.633e+00   0.054  0.95658    
## best_dir_winyes                 2.228e+00  2.316e+00   0.962  0.33670    
## top200_boxyes                   2.913e+00  3.697e+00   0.788  0.43114    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.97 on 416 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.851 
## F-statistic: 78.69 on 33 and 416 DF,  p-value: < 2.2e-16
```

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 "audience_score" has the highest p-value now, so we will remove the variable.

```
# without audience_score..

m3 <- lm(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 + best_pic_nom + best_pic_win + best_actor_win + 
           best_dir_win + top200_box, data= training_set)

summary(m3)

## 
## Call:
## lm(formula = 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 + 
##     best_pic_nom + best_pic_win + best_actor_win + best_dir_win + 
##     top200_box, data = training_set)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7097  -7.6050  -0.0239   6.9862  28.3659
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.234e+02  2.456e+02   0.910  0.36350
## title_typeFeature Film          -1.160e+01  4.708e+00  -2.465  0.01410 *
## title_typeTV Movie              -1.622e+00  7.414e+00  -0.219  0.82691
## genreAnimation                  -8.103e-02  5.815e+00  -0.014  0.98889
## genreArt House & International -4.709e+00  4.155e+00  -1.133  0.25768
## genreComedy                      1.343e+00  2.172e+00   0.618  0.53664
## genreDocumentary                -7.687e+00  5.014e+00  -1.533  0.12598
## genreDrama                       2.554e+00  1.949e+00   1.310  0.19079
## genreHorror                      1.799e+00  3.503e+00   0.513  0.60795
## genreMusical & Performing Arts -5.105e+00  6.285e+00  -0.812  0.41712
## genreMystery & Suspense         -7.418e-01  2.490e+00  -0.298  0.76593
## genreOther                       1.227e+00  3.990e+00   0.308  0.75858
## genreScience Fiction & Fantasy -1.588e+00  4.542e+00  -0.350  0.72674
## runtime                         -2.426e-02  3.712e-02  -0.654  0.51373
## mpaa_ratingNC-17                -1.726e+00  8.599e+00  -0.201  0.84097
## mpaa_ratingPG                   -2.025e+00  3.524e+00  -0.574  0.56599
## mpaa_ratingPG-13                -3.323e+00  3.708e+00  -0.896  0.37069
## mpaa_ratingR                    -3.334e+00  3.511e+00  -0.950  0.34281
## mpaa_ratingUnrated               6.252e-01  4.191e+00   0.149  0.88148
## thtr_rel_year                   -2.043e-01  7.280e-02  -2.807  0.00524 **
## thtr_rel_month                  -1.590e-01  1.562e-01  -1.018  0.30908
## dvd_rel_year                     1.108e-01  1.535e-01   0.722  0.47096
## dvd_rel_month                   -1.606e-01  1.606e-01  -1.000  0.31769
## imdb_rating                      9.377e+00  8.257e-01  11.357  < 2e-16 ***
## imdb_num_votes                  -1.340e-05  6.374e-06  -2.102  0.03613 *
## critics_ratingFresh             -8.613e+00  1.675e+00  -5.143 4.18e-07 ***
## critics_ratingRotten            -4.111e+01  1.840e+00 -22.339  < 2e-16 ***
## audience_ratingUpright          -7.974e-01  1.541e+00  -0.517  0.60510
## best_pic_nomyes                  2.086e+00  3.234e+00   0.645  0.51931
## best_pic_winyes                  5.462e-01  6.133e+00   0.089  0.92908
## best_actor_winyes                8.752e-02  1.630e+00   0.054  0.95722
## best_dir_winyes                  2.228e+00  2.313e+00   0.963  0.33610
## top200_boxyes                    2.910e+00  3.692e+00   0.788  0.43099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.96 on 417 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.8513
## F-statistic: 81.35 on 32 and 417 DF,  p-value: < 2.2e-16
```

Looking again at the model we will remove “best_actor_win” as it’s p-value is highest.

```
# without best_actor_win variable..

m4 <- lm(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 + best_pic_nom + best_pic_win +
           best_dir_win + top200_box, data= training_set)

summary(m4)

##
## Call:
## lm(formula = 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 +
##     best_pic_nom + best_pic_win + best_dir_win + top200_box,
##     data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.722  -7.610   0.007   7.000  28.358
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.243e+02  2.448e+02   0.917  0.35988
## title_typeFeature Film         -1.160e+01  4.701e+00  -2.467  0.01401 *
## title_typeTV Movie             -1.624e+00  7.405e+00  -0.219  0.82656
## genreAnimation                 -7.682e-02  5.808e+00  -0.013  0.98945
## genreArt House & International -4.714e+00  4.149e+00  -1.136  0.25645
## genreComedy                     1.341e+00  2.169e+00   0.618  0.53672
## genreDocumentary               -7.679e+00  5.005e+00  -1.534  0.12575
## genreDrama                      2.555e+00  1.947e+00   1.313  0.19002
## genreHorror                     1.795e+00  3.499e+00   0.513  0.60809
## genreMusical & Performing Arts -5.095e+00  6.275e+00  -0.812  0.41730
## genreMystery & Suspense        -7.331e-01  2.482e+00  -0.295  0.76785
## genreOther                      1.232e+00  3.985e+00   0.309  0.75733
## genreScience Fiction & Fantasy -1.591e+00  4.536e+00  -0.351  0.72601
## runtime                        -2.386e-02  3.632e-02  -0.657  0.51155
## mpaa_ratingNC-17               -1.682e+00  8.548e+00  -0.197  0.84414
## mpaa_ratingPG                  -2.014e+00  3.515e+00  -0.573  0.56694
## mpaa_ratingPG-13               -3.315e+00  3.700e+00  -0.896  0.37089
## mpaa_ratingR                   -3.329e+00  3.505e+00  -0.950  0.34275
## mpaa_ratingUnrated              6.226e-01  4.186e+00   0.149  0.88183
## thtr_rel_year                  -2.040e-01  7.243e-02  -2.816  0.00509 **
## thtr_rel_month                 -1.591e-01  1.560e-01  -1.020  0.30826
## dvd_rel_year                    1.100e-01  1.526e-01   0.721  0.47153
## dvd_rel_month                  -1.616e-01  1.594e-01  -1.014  0.31121
## imdb_rating                     9.378e+00  8.247e-01  11.371  < 2e-16 ***
## imdb_num_votes                 -1.341e-05  6.362e-06  -2.108  0.03561 *
## critics_ratingFresh            -8.607e+00  1.669e+00  -5.158 3.87e-07 ***
## critics_ratingRotten           -4.111e+01  1.838e+00 -22.367  < 2e-16 ***
## audience_ratingUpright         -8.026e-01  1.536e+00  -0.522  0.60163    
## best_pic_nomyes                 2.110e+00  3.199e+00   0.660  0.50993    
## best_pic_winyes                 5.447e-01  6.125e+00   0.089  0.92919    
## best_dir_winyes                 2.221e+00  2.307e+00   0.963  0.33627    
## top200_boxyes                   2.912e+00  3.687e+00   0.790  0.43003    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.95 on 418 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.8517 
## F-statistic: 84.17 on 31 and 418 DF,  p-value: < 2.2e-16
```

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

```
# removing best_pic_win variable

m5 <- lm(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 + best_pic_nom +
           best_dir_win + top200_box, data= training_set)

summary(m5)
```

```
## 
## Call:
## lm(formula = 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 + 
##     best_pic_nom + best_dir_win + top200_box, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.7262  -7.6077  -0.0003   6.9993  28.3508 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     2.238e+02  2.444e+02   0.916  0.36037    
## title_typeFeature Film         -1.161e+01  4.694e+00  -2.473  0.01380 *  
## title_typeTV Movie             -1.621e+00  7.397e+00  -0.219  0.82665    
## genreAnimation                 -7.131e-02  5.800e+00  -0.012  0.99020    
## genreArt House & International -4.699e+00  4.140e+00  -1.135  0.25701    
## genreComedy                     1.357e+00  2.159e+00   0.628  0.53003    
## genreDocumentary               -7.664e+00  4.997e+00  -1.534  0.12581    
## genreDrama                      2.561e+00  1.943e+00   1.318  0.18812    
## genreHorror                     1.801e+00  3.494e+00   0.515  0.60657    
## genreMusical & Performing Arts -5.103e+00  6.267e+00  -0.814  0.41591    
## genreMystery & Suspense        -7.273e-01  2.478e+00  -0.293  0.76931
## genreOther                      1.215e+00  3.975e+00   0.306  0.76011
## genreScience Fiction & Fantasy -1.591e+00  4.531e+00  -0.351  0.72569
## runtime                        -2.362e-02  3.618e-02  -0.653  0.51413
## mpaa_ratingNC-17               -1.680e+00  8.538e+00  -0.197  0.84406
## mpaa_ratingPG                  -2.020e+00  3.510e+00  -0.575  0.56536
## mpaa_ratingPG-13               -3.327e+00  3.693e+00  -0.901  0.36814
## mpaa_ratingR                   -3.334e+00  3.501e+00  -0.952  0.34146
## mpaa_ratingUnrated              6.229e-01  4.181e+00   0.149  0.88163
## thtr_rel_year                  -2.045e-01  7.207e-02  -2.838  0.00476 **
## thtr_rel_month                 -1.600e-01  1.555e-01  -1.029  0.30390
## dvd_rel_year                    1.108e-01  1.521e-01   0.728  0.46676
## dvd_rel_month                  -1.626e-01  1.588e-01  -1.024  0.30631
## imdb_rating                     9.375e+00  8.230e-01  11.391  < 2e-16 ***
## imdb_num_votes                 -1.330e-05  6.226e-06  -2.136  0.03325 *
## critics_ratingFresh            -8.614e+00  1.665e+00  -5.174 3.55e-07 ***
## critics_ratingRotten           -4.111e+01  1.836e+00 -22.393  < 2e-16 ***
## audience_ratingUpright         -8.029e-01  1.534e+00  -0.523  0.60103
## best_pic_nomyes                 2.206e+00  3.009e+00   0.733  0.46397
## best_dir_winyes                 2.286e+00  2.183e+00   1.047  0.29551
## top200_boxyes                   2.919e+00  3.682e+00   0.793  0.42823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 419 degrees of freedom
## Multiple R-squared:  0.8619, Adjusted R-squared:  0.852
## F-statistic: 87.19 on 30 and 419 DF,  p-value: < 2.2e-16
```

Every time we remove variable, our Adjusted R-squared value is increasing. Now we will remove “genre” variable as p-value is very high for all the levels of “genre” variable.

```
# without genre

m6 <- lm(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 + best_pic_nom +
           best_dir_win + top200_box, data= training_set)

summary(m6)
```

```
##
## Call:
## lm(formula = 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 +
##     best_pic_nom + best_dir_win + top200_box, data = training_set)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -27.4927  -7.3251   0.2216   7.2933  28.6565 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            2.424e+02  2.421e+02   1.001  0.31731    
## title_typeFeature Film -3.777e+00  2.441e+00  -1.547  0.12253    
## title_typeTV Movie     7.832e+00  5.919e+00   1.323  0.18645    
## runtime               -1.479e-02  3.439e-02  -0.430  0.66733    
## mpaa_ratingNC-17       7.093e-01  8.347e+00   0.085  0.93231    
## mpaa_ratingPG         -1.130e+00  3.163e+00  -0.357  0.72112    
## mpaa_ratingPG-13      -1.911e+00  3.263e+00  -0.586  0.55832    
## mpaa_ratingR          -2.097e+00  3.054e+00  -0.687  0.49266    
## mpaa_ratingUnrated     3.456e-01  3.832e+00   0.090  0.92818    
## thtr_rel_year         -1.982e-01  7.081e-02  -2.799  0.00536 ** 
## thtr_rel_month        -1.459e-01  1.542e-01  -0.947  0.34437    
## dvd_rel_year           9.114e-02  1.505e-01   0.606  0.54502    
## dvd_rel_month         -2.114e-01  1.573e-01  -1.344  0.17961    
## imdb_rating            9.371e+00  7.986e-01  11.734  < 2e-16 ***
## imdb_num_votes        -1.468e-05  6.089e-06  -2.411  0.01634 *  
## critics_ratingFresh   -8.921e+00  1.645e+00  -5.424 9.74e-08 ***
## critics_ratingRotten  -4.161e+01  1.810e+00 -22.990  < 2e-16 ***
## audience_ratingUpright -4.377e-01  1.507e+00  -0.290  0.77167    
## best_pic_nomyes        2.414e+00  3.005e+00   0.803  0.42216    
## best_dir_winyes        1.780e+00  2.168e+00   0.821  0.41224    
## top200_boxyes          3.191e+00  3.639e+00   0.877  0.38101    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.95 on 429 degrees of freedom
## Multiple R-squared:  0.8583, Adjusted R-squared:  0.8517 
## F-statistic: 129.9 on 20 and 429 DF,  p-value: < 2.2e-16
```

But we see that even though we removed genre variable our Adjusted R Squared value is reduced, so we will include genre in our model and now we remove "mpaa_rating" variable.

```
# without mpaa_rating

m7 <- lm(critics_score ~ title_type + genre + 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_dir_win + top200_box, data= training_set)

summary(m7)
```

```
## 
## Call:
## lm(formula = critics_score ~ title_type + genre + 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_dir_win + top200_box, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -28.0270  -7.5225   0.0371   7.2798  28.1784 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     1.875e+02  2.400e+02   0.781  0.43509    
## title_typeFeature Film         -1.212e+01  4.666e+00  -2.597  0.00972 ** 
## title_typeTV Movie             -5.433e-01  7.327e+00  -0.074  0.94093    
## genreAnimation                  2.090e+00  5.273e+00   0.396  0.69202    
## genreArt House & International -4.683e+00  4.071e+00  -1.150  0.25072    
## genreComedy                     9.534e-01  2.119e+00   0.450  0.65297    
## genreDocumentary               -6.741e+00  4.912e+00  -1.372  0.17068    
## genreDrama                      2.040e+00  1.885e+00   1.082  0.27980    
## genreHorror                     1.300e+00  3.378e+00   0.385  0.70049    
## genreMusical & Performing Arts -5.096e+00  6.230e+00  -0.818  0.41385    
## genreMystery & Suspense        -1.214e+00  2.428e+00  -0.500  0.61733    
## genreOther                      1.198e+00  3.937e+00   0.304  0.76107    
## genreScience Fiction & Fantasy -1.733e+00  4.496e+00  -0.385  0.70010    
## runtime                        -2.826e-02  3.509e-02  -0.805  0.42106    
## thtr_rel_year                  -2.221e-01  6.724e-02  -3.303  0.00104 ** 
## thtr_rel_month                 -1.592e-01  1.540e-01  -1.033  0.30202    
## dvd_rel_year                    1.453e-01  1.498e-01   0.970  0.33251    
## dvd_rel_month                  -1.882e-01  1.572e-01  -1.197  0.23209    
## imdb_rating                     9.494e+00  8.162e-01  11.632  < 2e-16 ***
## imdb_num_votes                 -1.434e-05  6.158e-06  -2.328  0.02037 *  
## critics_ratingFresh            -8.420e+00  1.649e+00  -5.106 4.99e-07 ***
## critics_ratingRotten           -4.107e+01  1.817e+00 -22.610  < 2e-16 ***
## audience_ratingUpright         -7.247e-01  1.522e+00  -0.476  0.63431    
## best_pic_nomyes                 2.166e+00  2.999e+00   0.722  0.47053    
## best_dir_winyes                 2.177e+00  2.174e+00   1.002  0.31700    
## top200_boxyes                   3.814e+00  3.601e+00   1.059  0.29009    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.91 on 424 degrees of freedom
## Multiple R-squared:  0.8609, Adjusted R-squared:  0.8527 
## F-statistic:   105 on 25 and 424 DF,  p-value: < 2.2e-16
```

Now we will remove “audience_rating” variable.

```
# without audience_rating variable.

m9 <- lm(critics_score ~ title_type + genre + runtime + thtr_rel_year + 
thtr_rel_month +
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
critics_rating +
           best_pic_nom +
           best_dir_win + top200_box, data= training_set)

summary(m9)
```

```
##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + thtr_rel_year
+
##     thtr_rel_month + dvd_rel_year + dvd_rel_month + imdb_rating +
##     imdb_num_votes + critics_rating + best_pic_nom + best_dir_win +
##     top200_box, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.6980  -7.3976   0.2198   7.2023  28.3626
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.770e+02  2.388e+02   0.741   0.4590
## title_typeFeature Film         -1.204e+01  4.659e+00  -2.584   0.0101 *
## title_typeTV Movie             -5.304e-01  7.320e+00  -0.072   0.9423
## genreAnimation                  1.864e+00  5.247e+00   0.355   0.7226
## genreArt House & International -4.782e+00  4.062e+00  -1.177   0.2397
## genreComedy                     9.278e-01  2.116e+00   0.438   0.6613
## genreDocumentary               -6.747e+00  4.908e+00  -1.375   0.1699
## genreDrama                      1.974e+00  1.878e+00   1.051   0.2938
## genreHorror                     1.391e+00  3.369e+00   0.413   0.6800
## genreMusical & Performing Arts -5.198e+00  6.221e+00  -0.836   0.4038
## genreMystery & Suspense        -1.173e+00  2.424e+00  -0.484   0.6288
## genreOther                      1.140e+00  3.932e+00   0.290   0.7720
## genreScience Fiction & Fantasy -1.660e+00  4.489e+00  -0.370   0.7117
## runtime                        -2.681e-02  3.493e-02  -0.768   0.4431
## thtr_rel_year                  -2.226e-01  6.717e-02  -3.314   0.0010 **
## thtr_rel_month                 -1.546e-01  1.536e-01  -1.006   0.3148
## dvd_rel_year                    1.513e-01  1.491e-01   1.015   0.3109
## dvd_rel_month                  -1.888e-01  1.571e-01  -1.202   0.2301
## imdb_rating                     9.307e+00  7.155e-01  13.008  < 2e-16 ***
## imdb_num_votes                 -1.441e-05  6.151e-06  -2.343   0.0196 *
## critics_ratingFresh            -8.348e+00  1.641e+00  -5.088 5.44e-07 ***
## critics_ratingRotten           -4.092e+01  1.787e+00 -22.897  < 2e-16 ***
## best_pic_nomyes                 2.117e+00  2.994e+00   0.707   0.4800
## best_dir_winyes                 2.213e+00  2.170e+00   1.020   0.3084
## top200_boxyes                   3.772e+00  3.596e+00   1.049   0.2949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 425 degrees of freedom
## Multiple R-squared:  0.8608, Adjusted R-squared:  0.853
## F-statistic: 109.5 on 24 and 425 DF,  p-value: < 2.2e-16
```

Now we will remove “best_pic_nom”

```
# Without best_pic_nom variable

m10 <- lm(critics_score ~ title_type + genre + runtime + thtr_rel_year +
thtr_rel_month +
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
critics_rating +
           best_dir_win + top200_box, data= training_set)

summary(m10)

##
## Call:
## lm(formula = critics_score ~ title_type + genre + runtime + thtr_rel_year
+
##     thtr_rel_month + dvd_rel_year + dvd_rel_month + imdb_rating +
##     imdb_num_votes + critics_rating + best_dir_win + top200_box,
##     data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.6881  -7.6329   0.3879   7.1477  28.4657
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.750e+02  2.386e+02   0.734 0.463653
## title_typeFeature Film         -1.198e+01  4.655e+00  -2.573 0.010434 *
## title_typeTV Movie             -4.235e-01  7.314e+00  -0.058 0.953852
## genreAnimation                  1.893e+00  5.244e+00   0.361 0.718262
## genreArt House & International -4.757e+00  4.059e+00  -1.172 0.241915
## genreComedy                     1.007e+00  2.112e+00   0.477 0.633731
## genreDocumentary               -6.658e+00  4.903e+00  -1.358 0.175259
## genreDrama                      2.037e+00  1.875e+00   1.086 0.277978
## genreHorror                     1.480e+00  3.365e+00   0.440 0.660303
## genreMusical & Performing Arts -5.254e+00  6.217e+00  -0.845 0.398487
## genreMystery & Suspense        -1.121e+00  2.421e+00  -0.463 0.643581
## genreOther                      1.261e+00  3.926e+00   0.321 0.748245
## genreScience Fiction & Fantasy -1.622e+00  4.486e+00  -0.362 0.717810
## runtime                        -2.275e-02  3.443e-02  -0.661 0.509141
## thtr_rel_year                  -2.267e-01  6.688e-02  -3.390 0.000763 ***
## thtr_rel_month                 -1.439e-01  1.528e-01  -0.942 0.346764
## dvd_rel_year                    1.561e-01  1.489e-01   1.049 0.294956
## dvd_rel_month                  -1.875e-01  1.570e-01  -1.194 0.233044
## imdb_rating                     9.328e+00  7.145e-01  13.055  < 2e-16 ***
## imdb_num_votes                 -1.385e-05  6.095e-06  -2.272 0.023605 *
## critics_ratingFresh            -8.505e+00  1.625e+00  -5.235 2.59e-07 ***
## critics_ratingRotten           -4.107e+01  1.774e+00 -23.156  < 2e-16 ***
## best_dir_winyes                 2.216e+00  2.169e+00   1.021 0.307623    
## top200_boxyes                   3.792e+00  3.594e+00   1.055 0.292029    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.89 on 426 degrees of freedom
## Multiple R-squared:  0.8607, Adjusted R-squared:  0.8531 
## F-statistic: 114.4 on 23 and 426 DF,  p-value: < 2.2e-16
```

Removing “runtime” variable..

```
# removing runtime variable

m11 <- lm(critics_score ~ title_type + genre + thtr_rel_year + thtr_rel_month
+
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
critics_rating +
           best_dir_win + top200_box, data= training_set)

summary(m11)
```

```
## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     thtr_rel_month + dvd_rel_year + dvd_rel_month + imdb_rating + 
##     imdb_num_votes + critics_rating + best_dir_win + top200_box, 
##     data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.7438  -7.6476   0.2496   6.9972  28.3961 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     1.601e+02  2.374e+02   0.675 0.500328    
## title_typeFeature Film         -1.195e+01  4.652e+00  -2.568 0.010555 *  
## title_typeTV Movie             -3.375e-01  7.308e+00  -0.046 0.963188    
## genreAnimation                  2.293e+00  5.205e+00   0.440 0.659855    
## genreArt House & International -4.746e+00  4.057e+00  -1.170 0.242654    
## genreComedy                     1.106e+00  2.105e+00   0.525 0.599673    
## genreDocumentary               -6.448e+00  4.890e+00  -1.319 0.188015    
## genreDrama                      1.899e+00  1.862e+00   1.020 0.308411    
## genreHorror                     1.732e+00  3.341e+00   0.518 0.604420    
## genreMusical & Performing Arts -5.522e+00  6.199e+00  -0.891 0.373579    
## genreMystery & Suspense        -1.284e+00  2.407e+00  -0.533 0.594061    
## genreOther                      1.253e+00  3.923e+00   0.319 0.749552    
## genreScience Fiction & Fantasy -1.401e+00  4.471e+00  -0.313 0.754116    
## thtr_rel_year                  -2.229e-01  6.658e-02  -3.348 0.000886 ***
## thtr_rel_month                 -1.618e-01  1.502e-01  -1.077 0.282049    
## dvd_rel_year                     1.588e-01  1.487e-01   1.068 0.286179    
## dvd_rel_month                   -1.856e-01  1.569e-01  -1.183 0.237427    
## imdb_rating                      9.273e+00  7.092e-01  13.076  < 2e-16 ***
## imdb_num_votes                  -1.476e-05  5.931e-06  -2.489 0.013188 *  
## critics_ratingFresh             -8.479e+00  1.623e+00  -5.224 2.74e-07 ***
## critics_ratingRotten            -4.113e+01  1.770e+00 -23.234  < 2e-16 ***
## best_dir_winyes                  1.900e+00  2.114e+00   0.899 0.369343    
## top200_boxyes                    3.741e+00  3.591e+00   1.042 0.298111    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.89 on 427 degrees of freedom
## Multiple R-squared:  0.8605, Adjusted R-squared:  0.8533 
## F-statistic: 119.7 on 22 and 427 DF,  p-value: < 2.2e-16
```

Now we will remove “thtr_rel_month”.

```
# removing thtr_rel_month
m12 <- lm(critics_score ~ title_type + genre + thtr_rel_year +
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
critics_rating +
           best_dir_win + top200_box, data= training_set)

summary(m12)

## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes + 
##     critics_rating + best_dir_win + top200_box, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.7841  -7.4540  -0.1204   7.1407  29.3323 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)   
## (Intercept)                     1.620e+02  2.374e+02   0.682  0.49539   
## title_typeFeature Film         -1.206e+01  4.652e+00  -2.593  0.00983 **
## title_typeTV Movie             -1.390e-01  7.307e+00  -0.019  0.98483   
## genreAnimation                  2.432e+00  5.205e+00   0.467  0.64059   
## genreArt House & International -4.793e+00  4.057e+00  -1.181  0.23816   
## genreComedy                     9.846e-01  2.103e+00   0.468  0.63983   
## genreDocumentary               -6.473e+00  4.891e+00  -1.324  0.18636   
## genreDrama                      1.809e+00  1.860e+00   0.972  0.33147   
## genreHorror                     1.716e+00  3.342e+00   0.514  0.60784   
## genreMusical & Performing Arts -6.073e+00  6.179e+00  -0.983  0.32623   
## genreMystery & Suspense        -1.274e+00  2.408e+00  -0.529  0.59695   
## genreOther                      1.363e+00  3.923e+00   0.347  0.72848   
## genreScience Fiction & Fantasy -1.326e+00  4.471e+00  -0.296  0.76700   
## thtr_rel_year                  -2.231e-01  6.660e-02  -3.350  0.00088 ***
## dvd_rel_year                    1.578e-01  1.488e-01   1.061  0.28940    
## dvd_rel_month                  -1.609e-01  1.552e-01  -1.037  0.30045    
## imdb_rating                     9.190e+00  7.051e-01  13.033  < 2e-16 ***
## imdb_num_votes                 -1.488e-05  5.931e-06  -2.509  0.01247 *  
## critics_ratingFresh            -8.370e+00  1.620e+00  -5.166 3.67e-07 ***
## critics_ratingRotten           -4.116e+01  1.770e+00 -23.252  < 2e-16 ***
## best_dir_winyes                 1.795e+00  2.113e+00   0.850  0.39587    
## top200_boxyes                   3.526e+00  3.586e+00   0.983  0.32604    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.89 on 428 degrees of freedom
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8533 
## F-statistic: 125.3 on 21 and 428 DF,  p-value: < 2.2e-16
```

We see that removing “thtr_rel_month” our Adjusted R squared value is same but we will not include “thtr_rel_month” variable as it’s p-value is very high so it does not add anything useful in our model. Now Removing “best_director_win” variable.

```
m13 <- lm(critics_score ~ title_type + genre + thtr_rel_year +
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes +
critics_rating +
           top200_box, data= training_set)

summary(m13)
```

```
## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes + 
##     critics_rating + top200_box, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.8399  -7.5767   0.3916   7.0776  29.3205 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     1.761e+02  2.368e+02   0.744 0.457454    
## title_typeFeature Film         -1.183e+01  4.642e+00  -2.549 0.011140 *  
## title_typeTV Movie              5.063e-02  7.302e+00   0.007 0.994470    
## genreAnimation                  2.231e+00  5.198e+00   0.429 0.667917    
## genreArt House & International -4.943e+00  4.052e+00  -1.220 0.223225    
## genreComedy                     9.332e-01  2.101e+00   0.444 0.657155    
## genreDocumentary               -6.415e+00  4.889e+00  -1.312 0.190126    
## genreDrama                      1.739e+00  1.858e+00   0.936 0.349822    
## genreHorror                     1.719e+00  3.341e+00   0.514 0.607197    
## genreMusical & Performing Arts -5.744e+00  6.165e+00  -0.932 0.351993    
## genreMystery & Suspense        -1.253e+00  2.407e+00  -0.521 0.602952    
## genreOther                      1.248e+00  3.919e+00   0.319 0.750246    
## genreScience Fiction & Fantasy -1.277e+00  4.470e+00  -0.286 0.775210    
## thtr_rel_year                  -2.286e-01  6.626e-02  -3.450 0.000616 ***
## dvd_rel_year                    1.562e-01  1.487e-01   1.051 0.293988    
## dvd_rel_month                  -1.701e-01  1.548e-01  -1.099 0.272265    
## imdb_rating                     9.200e+00  7.048e-01  13.053  < 2e-16 ***
## imdb_num_votes                 -1.402e-05  5.842e-06  -2.400 0.016813 *  
## critics_ratingFresh            -8.367e+00  1.620e+00  -5.166 3.68e-07 ***
## critics_ratingRotten           -4.128e+01  1.764e+00 -23.398  < 2e-16 ***
## top200_boxyes                   3.322e+00  3.577e+00   0.929 0.353499    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.88 on 429 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.8534 
## F-statistic: 131.7 on 20 and 429 DF,  p-value: < 2.2e-16
```

Now removing “top200_box” variable.

```
m14 <- lm(critics_score ~ title_type + genre + thtr_rel_year + 
           dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes + 
critics_rating, data= training_set)

summary(m14)
```

```
## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     dvd_rel_year + dvd_rel_month + imdb_rating + imdb_num_votes + 
##     critics_rating, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.8821  -7.6207   0.5355   7.0867  29.3528 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     1.781e+02  2.367e+02   0.752 0.452290    
## title_typeFeature Film         -1.185e+01  4.642e+00  -2.554 0.011007 *  
## title_typeTV Movie              2.620e-02  7.300e+00   0.004 0.997138    
## genreAnimation                  1.924e+00  5.186e+00   0.371 0.710800    
## genreArt House & International -5.142e+00  4.046e+00  -1.271 0.204471    
## genreComedy                     7.495e-01  2.092e+00   0.358 0.720241    
## genreDocumentary               -6.577e+00  4.885e+00  -1.347 0.178845    
## genreDrama                      1.555e+00  1.847e+00   0.842 0.400448    
## genreHorror                     1.486e+00  3.331e+00   0.446 0.655752    
## genreMusical & Performing Arts -5.943e+00  6.160e+00  -0.965 0.335209    
## genreMystery & Suspense        -1.543e+00  2.386e+00  -0.647 0.518172    
## genreOther                      1.223e+00  3.918e+00   0.312 0.755026    
## genreScience Fiction & Fantasy -1.609e+00  4.454e+00  -0.361 0.718090    
## thtr_rel_year                    -2.376e-01  6.554e-02  -3.625 0.000323 ***
## dvd_rel_year                      1.644e-01  1.484e-01   1.108 0.268621    
## dvd_rel_month                    -1.593e-01  1.543e-01  -1.032 0.302469    
## imdb_rating                       9.176e+00  7.042e-01  13.030  < 2e-16 ***
## imdb_num_votes                   -1.271e-05  5.668e-06  -2.242 0.025442 *  
## critics_ratingFresh              -8.499e+00  1.613e+00  -5.268 2.18e-07 ***
## critics_ratingRotten             -4.139e+01  1.760e+00 -23.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.88 on 430 degrees of freedom
## Multiple R-squared:  0.8596, Adjusted R-squared:  0.8534 
## F-statistic: 138.6 on 19 and 430 DF,  p-value: < 2.2e-16
```

Now removing “dvd_rel_month” variable..

```
# removing dvd_rel_month variable..

m15 <- lm(critics_score ~ title_type + genre + thtr_rel_year + 
           dvd_rel_year + imdb_rating + imdb_num_votes + critics_rating,
data= training_set)

summary(m15)

## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     dvd_rel_year + imdb_rating + imdb_num_votes + critics_rating, 
##     data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26.9113  -7.7923   0.1912   7.1577  29.3816 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                       1.604e+02  2.361e+02   0.679 0.497365    
## title_typeFeature Film           -1.197e+01  4.641e+00  -2.578 0.010257 *  
## title_typeTV Movie               -1.913e-01  7.298e+00  -0.026 0.979103    
## genreAnimation                    2.036e+00  5.186e+00   0.393 0.694822    
## genreArt House & International -5.170e+00  4.046e+00  -1.278 0.202006    
## genreComedy                       8.372e-01  2.090e+00   0.401 0.688922    
## genreDocumentary                 -6.759e+00  4.882e+00  -1.385 0.166896    
## genreDrama                        1.620e+00  1.846e+00   0.877 0.380859    
## genreHorror                       1.526e+00  3.331e+00   0.458 0.646979    
## genreMusical & Performing Arts -5.728e+00  6.157e+00  -0.930 0.352751    
## genreMystery & Suspense          -1.567e+00  2.386e+00  -0.657 0.511771    
## genreOther                        1.492e+00  3.910e+00   0.382 0.702926    
## genreScience Fiction & Fantasy -1.932e+00  4.444e+00  -0.435 0.663995    
## thtr_rel_year                    -2.404e-01  6.548e-02  -3.671 0.000272 ***
## dvd_rel_year                    1.757e-01  1.480e-01   1.187 0.235832    
## imdb_rating                     9.132e+00  7.030e-01  12.990  < 2e-16 ***
## imdb_num_votes                 -1.284e-05  5.667e-06  -2.267 0.023899 *  
## critics_ratingFresh            -8.496e+00  1.613e+00  -5.266  2.2e-07 ***
## critics_ratingRotten           -4.143e+01  1.760e+00 -23.541  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.88 on 431 degrees of freedom
## Multiple R-squared:  0.8593, Adjusted R-squared:  0.8534 
## F-statistic: 146.2 on 18 and 431 DF,  p-value: < 2.2e-16
```

Removing “dvd_rel_year”..

```
m16 <- lm(critics_score ~ title_type + genre + thtr_rel_year +
           imdb_rating + imdb_num_votes + critics_rating, data= training_set)

summary(m16)
```

```
## 
## Call:
## lm(formula = critics_score ~ title_type + genre + thtr_rel_year + 
##     imdb_rating + imdb_num_votes + critics_rating, data = training_set)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.1033  -7.9721  -0.0962   7.2871  29.0644 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     4.135e+02  1.014e+02   4.077 5.43e-05 ***
## title_typeFeature Film         -1.203e+01  4.643e+00  -2.592 0.009861 ** 
## title_typeTV Movie             -5.061e-01  7.297e+00  -0.069 0.944738    
## genreAnimation                  1.939e+00  5.188e+00   0.374 0.708729    
## genreArt House & International -4.626e+00  4.022e+00  -1.150 0.250716    
## genreComedy                     8.106e-01  2.091e+00   0.388 0.698434    
## genreDocumentary               -6.508e+00  4.880e+00  -1.334 0.183006    
## genreDrama                      1.719e+00  1.845e+00   0.932 0.352002    
## genreHorror                     1.743e+00  3.327e+00   0.524 0.600762    
## genreMusical & Performing Arts -5.409e+00  6.154e+00  -0.879 0.379920    
## genreMystery & Suspense        -1.481e+00  2.386e+00  -0.621 0.535063    
## genreOther                      1.809e+00  3.903e+00   0.463 0.643288    
## genreScience Fiction & Fantasy -1.722e+00  4.442e+00  -0.388 0.698436    
## thtr_rel_year                  -1.905e-01  5.025e-02  -3.792 0.000171 ***
## imdb_rating                     9.041e+00  6.991e-01  12.932  < 2e-16 ***
## imdb_num_votes                 -1.252e-05  5.663e-06  -2.211 0.027550 *  
## critics_ratingFresh            -8.434e+00  1.613e+00  -5.228 2.68e-07 ***
## critics_ratingRotten           -4.158e+01  1.756e+00 -23.678  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.89 on 432 degrees of freedom
## Multiple R-squared:  0.8588, Adjusted R-squared:  0.8533
## F-statistic: 154.6 on 17 and 432 DF,  p-value: < 2.2e-16
```

We see that our Adjusted R squared values decreses by small value so we will include “dvd_rel_year” variable. Now we remove “title_type” variable..

```
m17 <- lm(critics_score ~ genre + thtr_rel_year +
           imdb_rating + imdb_num_votes + critics_rating, data= training_set)

summary(m17)
```

```
##
## Call:
## lm(formula = critics_score ~ genre + thtr_rel_year + imdb_rating +
##     imdb_num_votes + critics_rating, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2575  -7.7383   0.0035   7.2780  28.7233
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.773e+02  1.018e+02   3.704 0.000239 ***
## genreAnimation                  1.763e+00  5.241e+00   0.336 0.736790
## genreArt House & International -4.879e+00  4.062e+00  -1.201 0.230384
## genreComedy                     1.108e+00  2.109e+00   0.526 0.599480
## genreDocumentary                3.974e+00  2.807e+00   1.416 0.157568
## genreDrama                      1.684e+00  1.859e+00   0.906 0.365523
## genreHorror                     1.702e+00  3.362e+00   0.506 0.612835
## genreMusical & Performing Arts  3.757e-01  5.803e+00   0.065 0.948417
## genreMystery & Suspense        -1.576e+00  2.410e+00  -0.654 0.513584
## genreOther                      2.872e+00  3.896e+00   0.737 0.461459
## genreScience Fiction & Fantasy -1.762e+00  4.489e+00  -0.392 0.694905
## thtr_rel_year                  -1.785e-01  5.064e-02  -3.524 0.000470 ***
## imdb_rating                     9.127e+00  7.005e-01  13.029  < 2e-16 ***
## imdb_num_votes                 -1.424e-05  5.690e-06  -2.503 0.012689 *
## critics_ratingFresh            -8.534e+00  1.626e+00  -5.248 2.41e-07 ***
## critics_ratingRotten           -4.192e+01  1.768e+00 -23.706  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11 on 434 degrees of freedom
## Multiple R-squared:  0.8552, Adjusted R-squared:  0.8502
## F-statistic: 170.9 on 15 and 434 DF,  p-value: < 2.2e-16
```

We see that adjusted R -squared value is decreased by a large amount, so we will include “title_type” variable in our model. Now we see that there are no more variable we can remove as p-value of rest variable is less than 0.05, and by removing variables with p value greater than 0.05 our Adjusted R squared value decreases so we cannot remove them. So this is the final model.

```
# Final Model

m18 <- lm(critics_score ~ genre + title_type + thtr_rel_year +
           imdb_rating + imdb_num_votes + critics_rating + dvd_rel_year,
data= training_set)

summary(m18)

##
## Call:
## lm(formula = critics_score ~ genre + title_type + thtr_rel_year +
##     imdb_rating + imdb_num_votes + critics_rating + dvd_rel_year,
##     data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.9113  -7.7923   0.1912   7.1577  29.3816
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.604e+02  2.361e+02   0.679 0.497365
## genreAnimation                  2.036e+00  5.186e+00   0.393 0.694822
## genreArt House & International -5.170e+00  4.046e+00  -1.278 0.202006
## genreComedy                     8.372e-01  2.090e+00   0.401 0.688922
## genreDocumentary               -6.759e+00  4.882e+00  -1.385 0.166896
## genreDrama                      1.620e+00  1.846e+00   0.877 0.380859
## genreHorror                     1.526e+00  3.331e+00   0.458 0.646979
## genreMusical & Performing Arts -5.728e+00  6.157e+00  -0.930 0.352751
## genreMystery & Suspense        -1.567e+00  2.386e+00  -0.657 0.511771
## genreOther                      1.492e+00  3.910e+00   0.382 0.702926
## genreScience Fiction & Fantasy -1.932e+00  4.444e+00  -0.435 0.663995
## title_typeFeature Film         -1.197e+01  4.641e+00  -2.578 0.010257 *
## title_typeTV Movie             -1.913e-01  7.298e+00  -0.026 0.979103
## thtr_rel_year                  -2.404e-01  6.548e-02  -3.671 0.000272 ***
## imdb_rating                     9.132e+00  7.030e-01  12.990  < 2e-16 ***
## imdb_num_votes                 -1.284e-05  5.667e-06  -2.267 0.023899 *
## critics_ratingFresh            -8.496e+00  1.613e+00  -5.266  2.2e-07 ***
## critics_ratingRotten           -4.143e+01  1.760e+00 -23.541  < 2e-16 ***
## dvd_rel_year                    1.757e-01  1.480e-01   1.187 0.235832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.88 on 431 degrees of freedom
## Multiple R-squared:  0.8593, Adjusted R-squared:  0.8534
## F-statistic: 146.2 on 18 and 431 DF,  p-value: < 2.2e-16
```

Interpretations of model coefficients:

Consider the intercept in the final model. This shows that if a “Certified Fresh”, “action and adventure” movie with title type of “TV Movie” doesn’t mention it’s year of release , number of voters in IMDB, IMDB rating and it’s dvd release year then it’s critic’s rating is likely to be 1.604e+02.

For categorical variables like “genre”, “title_type” and “critics_rating”, one level of the variable is kept 0 which means that the level which is made 0, doesn’t affect the critic rating of the movie. And the interpretation of other levels is made by keeping one level 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 2.404e-01.

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 1.284e-05.

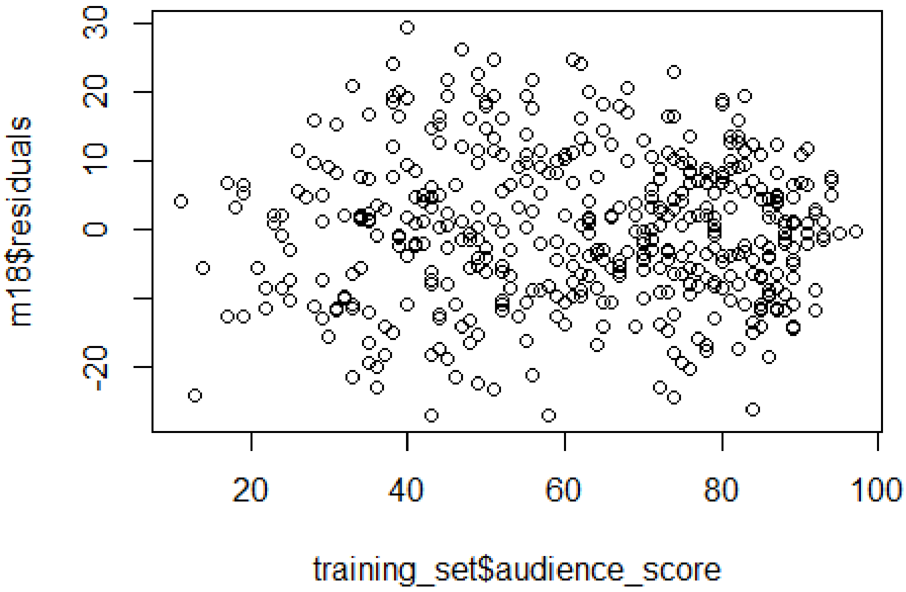
For ex. keeping other variables constant we can say that, on average, animation movies have critic’s rating greater than Action and Adventure by 2.036e+00.

Model Diagnostics:

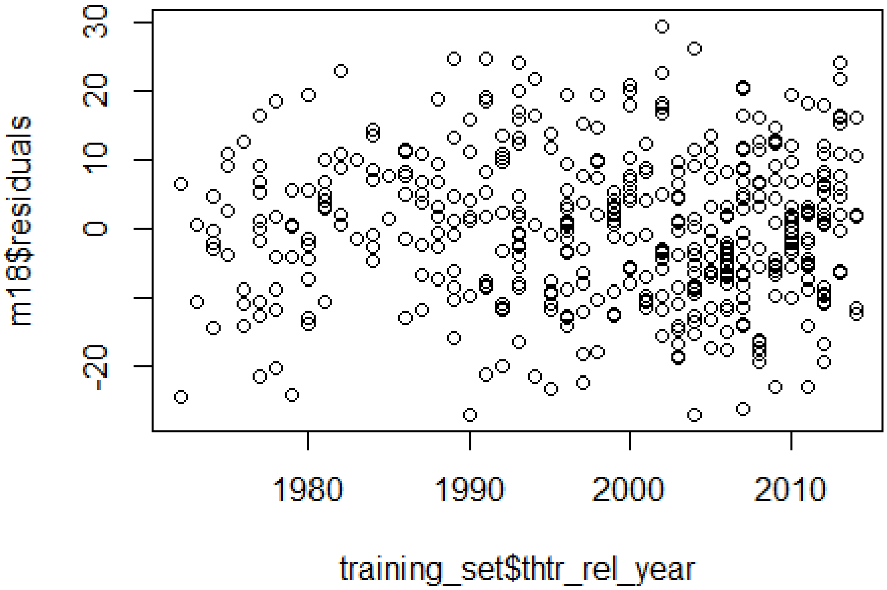
So we made our model. Now we check conditions required for multiple regression to be mapped valid.

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(m18$residuals ~ training_set$audience_score)
```



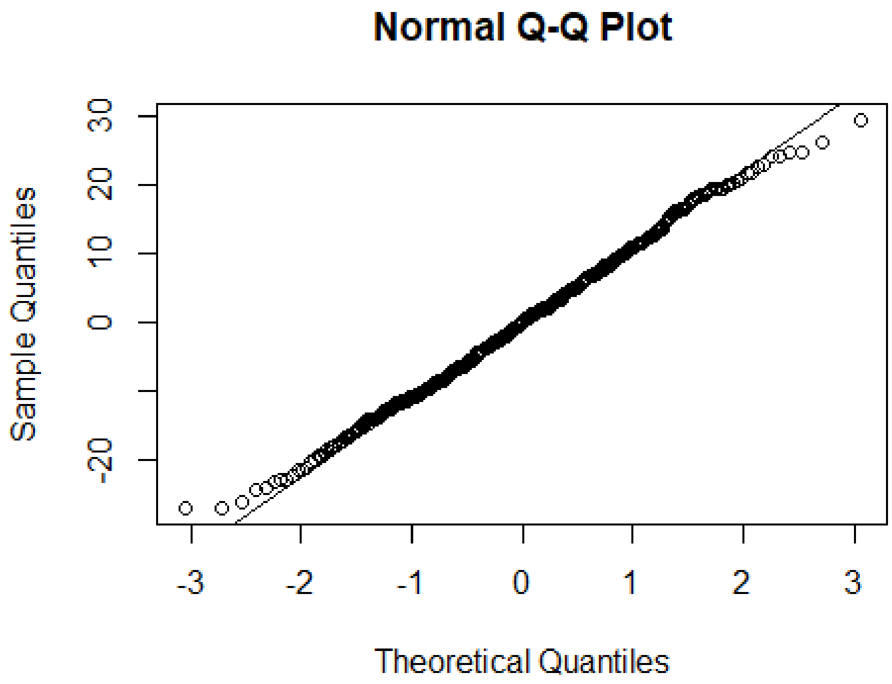
```
plot(m18$residuals ~ training_set$thtr_rel_year)
```



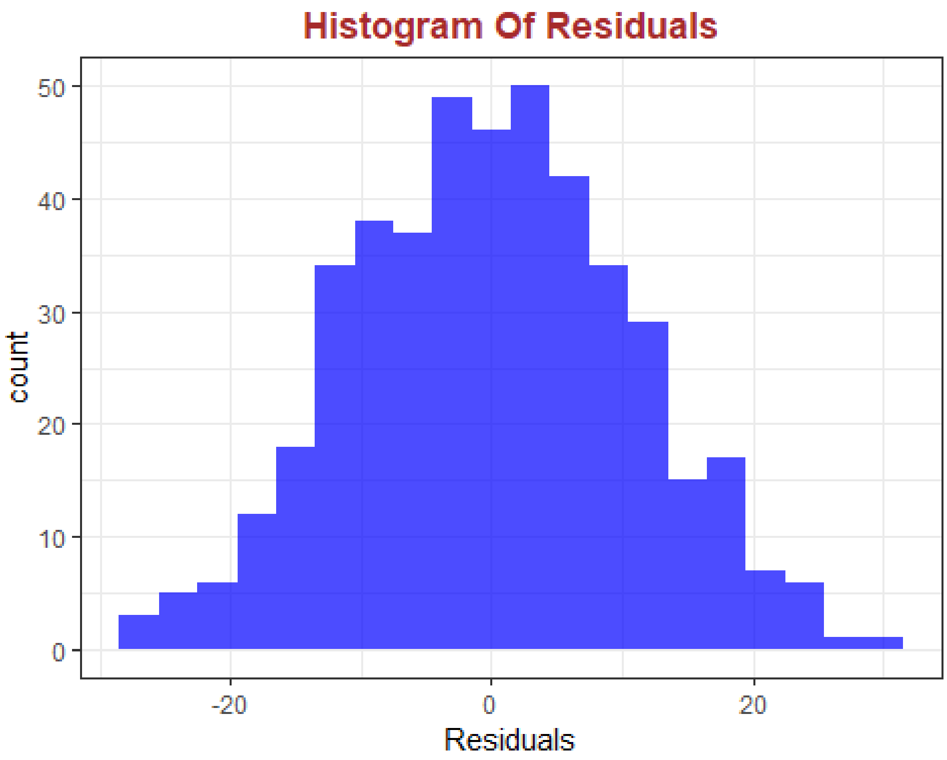
So we can see there is random scatter around 0.

2. The second condition for model diagnstics is nearly normal residual with mean 0.We can check it through histogram or we can see ot through normal probability plot.

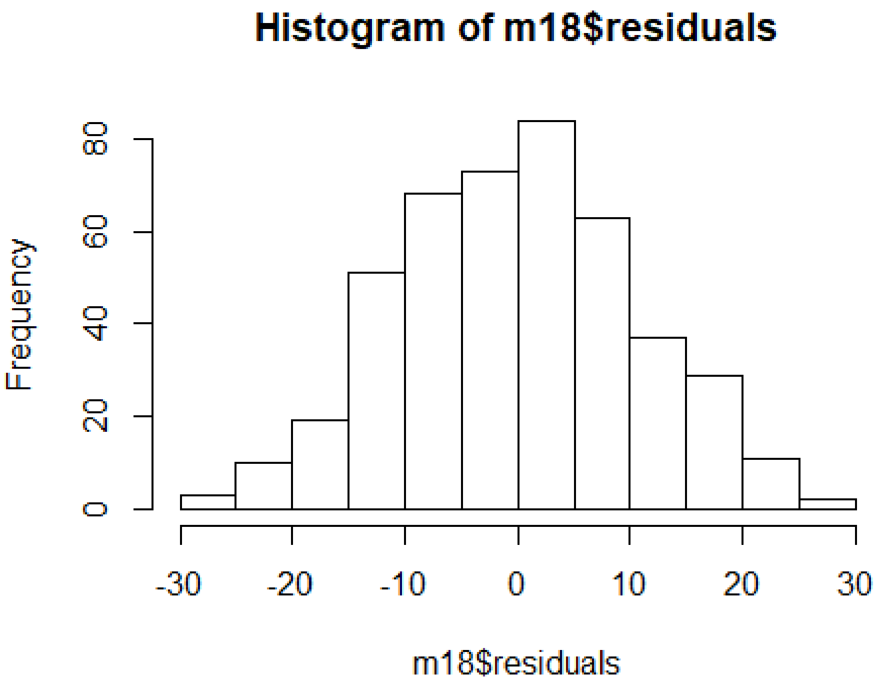
```
# plot with normal probability plot.
qqnorm(m18$residuals)
qqline(m18$residuals)
```



```
# Grab residuals
res <- residuals(m18)

# Convert to DataFrame for gglpot
res <- as.data.frame(res)

head(res)

##          res
## 1 16.4112699
## 2  8.9180144
## 3  0.8884743
## 4 -7.7990330
## 5  4.4497458
## 6  8.2065272

# Histogram of residuals
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"))
```



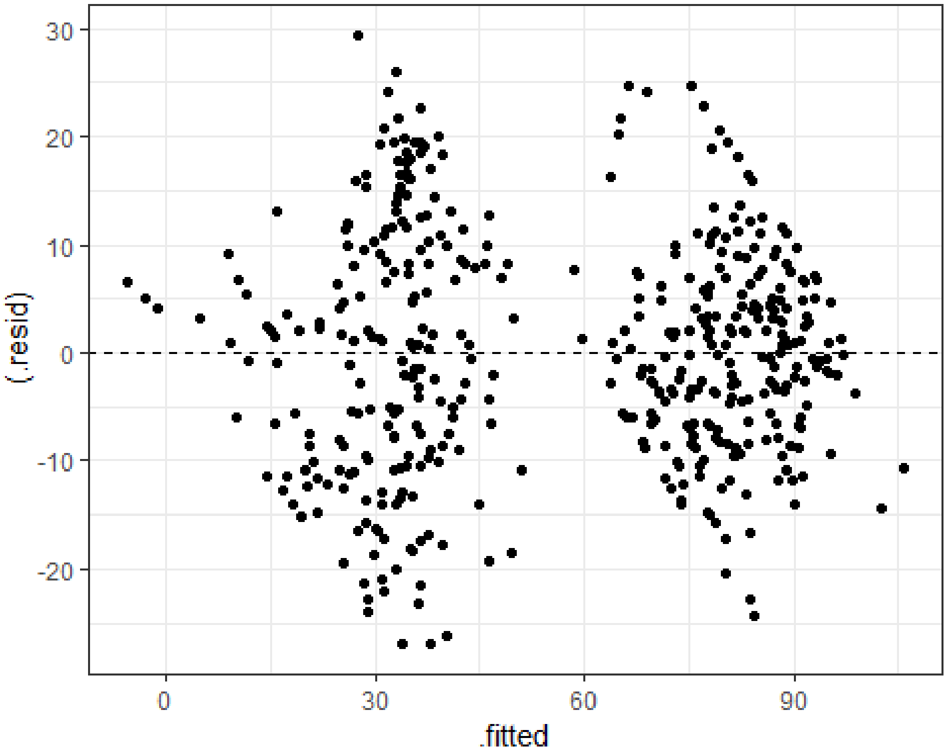
```
# or you can do the same with
hist(m18$residuals)
```



We can see that the distribution of residuals is nearly normal.

3. The third condition is constant variability of the residuals. We can do this by checking residuals plots vs predicted value.

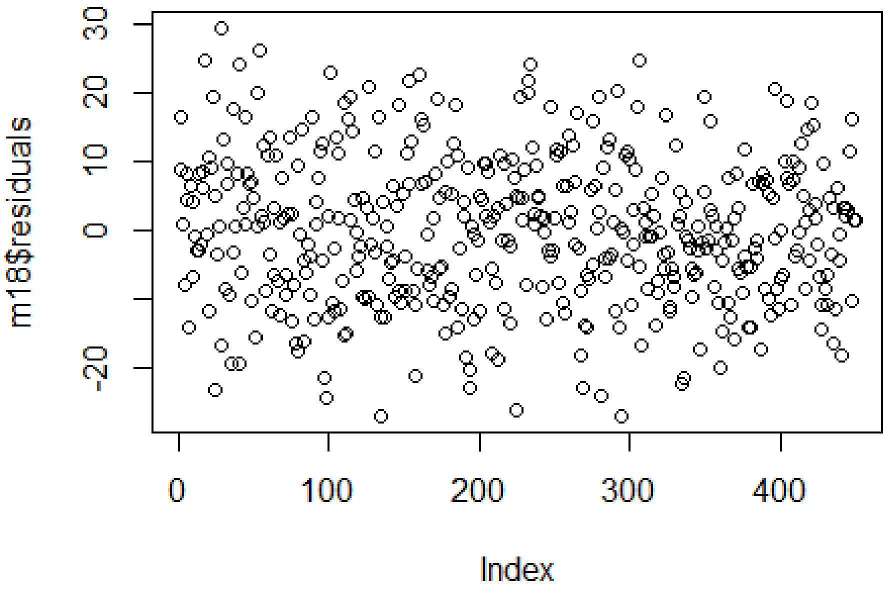
```
ggplot(data = m18, aes(x = .fitted, y = (.resid))) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") + theme_bw()
```



We can say to an extent that there is constant variability in the residuals.

4. The last diagnostic method is independent residuals.The independent residuals comes from independent observations.

```
plot(m18$residuals)
```



By looking at the plot there is random scatter about 0 and therefore we can say that the residuals are independent .

Part 5: Prediction

Now we see how our model predicted the test data, we will use predict() function for this purpose.

```
model.predictions <- predict(m18,  test_set)
results <- cbind(model.predictions,test_set$critics_score)
colnames(results) <- c('pred','real')
results <- as.data.frame(results)
head(results,10)
```

```
##        pred real
## 1  25.58180   33
## 2  79.97552   83
## 3  69.26353   67
## 4  86.54037   80
## 5  75.31488   61
## 6  19.04020   19
## 7  11.54072   29
## 8  85.20127   92
## 9  38.05257   47
## 10 93.16580   79
```

So we can see the predicted and original value, we can see that our model is not that bad, we made some great predictions, like for example in results data frame see 3rd observation, predicted is 69.263, and actual is 67. Also see 6th observation, real critic score is 19 and our model predicted 19.04 which is same, so great our model is good.

```
# Maximum value of prediction
max(results$pred)

## [1] 102.0408
```

But when we look at the maximum value of our prediction, we see that it is 102.0408, 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.

```
greater_hundred <- function(x){
    if  (x > 100){
        return(100)
    }else{
        return(x)
    }
}

results$pred <- sapply(results$pred,greater_hundred)
```

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

Information regarding audience_score, genre can be found at https://www.rottentomatoes.com/m/deadpool/

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

Information regarding genre can be found on https://www.imdb.com/title/tt1431045/?ref_=tt_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( genre= "Action & Adventure",thtr_rel_year = 2016,
title_type = "Feature Film", imdb_rating = 8, imdb_num_votes = 747563,
critics_rating = "Certified Fresh", dvd_rel_year = 2016)
```

Now let’s see what our model predicts.

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

predict(m18, newmovie)

##        1
## 81.43938
```

So our predicted value for rotten tomatoes critics rating is 81.43938 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(m18, newmovie, interval = "prediction", level = 0.95)

##        fit      lwr      upr
## 1 81.43938 58.82663 104.0521
```

The number 104.0521 is mere a number for confidence interval and as we know that critic score cannot exceed 100 so we can restrict our upper limit of confidence to 100. The above statement says that “We are 95% confident that the movie”DEADPOOL" will get critics score on average between 58.82663 and 100 by Rotten tomatoes."

Part 6: Conclusion

After making model and doing prediction, we can predict critic review of any movie given the specific parameters which are required for our model.

Thus we can find the factors which decides the success of the movie,like Imdb Rating, genre 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.