Data501 - Movies Project

Manjushree Raghwani, Niroopa Kannan, Varun Myla & Hashavardhan Bellala

Due Date = 12-16-2022

Research Question:

Q1) Oscar win actor/actress or director influence the overall performance or rating of the movie or not? OR Is an Oscar winning actor or actress in the cast associated with a better rating (in the context of multiple regression)?

Q2) Are Critics and Audience ratings close to each other or not. if not, are the critics ratings biased? OR Is critics' rating associated with audience's rating, in the context of regression model with other potential predictors.

Q3) Create additional variable 'Oscar' with a class "yes or no" to identify whether an movie has won atleast one Oscar award and create a model with other variable. Check if there is any significant difference after adding 'Oscar' in the model with other variable or not?

Q4) Create additional variable 'movie_score' which is a combination of imdb_rating & audience_score to check if they add more weightage to the response variable 'imdb_rating'?

Q5) Insights about movies characteristics with reference to title_type, mpaa_rating or genre.

Q6) Predicting imdb_rating. Show sample test case.

Importing Data set & Packages

```
library(rlang)
library(dplyr)
```

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

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

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.2.2
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: ’VIM’
```

```
## The following object is masked from ’package:datasets’:
##
##     sleep
```

```
library(psych)
```

```
## Warning: package ’psych’ was built under R version 4.2.2
```

```
##
## Attaching package: ’psych’
```

```
## The following object is masked from ’package:car’:
##
##     logit
```

```
## The following objects are masked from ’package:ggplot2’:
##
##     %+%, alpha
```

```
library(gridExtra)
```

```
##
## Attaching package: ’gridExtra’
```

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

```
library(xtable)
```

```
## Warning: package ’xtable’ was built under R version 4.2.2
```

```
library(faraway)
```

```
## Warning: package ’faraway’ was built under R version 4.2.2
```

```
##
## Attaching package: ’faraway’
```

```
## The following object is masked from ’package:psych’:
##
##     logit
```

```
## The following object is masked from ’package:VIM’:
##
##     diabetes
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
## The following object is masked from 'package:GGally':
##
##     happy
```

```
load(url("http://people.math.binghamton.edu/qiao/data501/data/movies.RData"))
```

```
orgdf <-movies
orgdf
```

```
## # A tibble: 651 x 32
##    title    title~1 genre runtime mpaa_~2 studio thtr_~3 thtr_~4 thtr_~5 dvd_r~6
##    <chr>    <fct>   <fct>   <dbl> <fct>   <fct>    <dbl>   <dbl>   <dbl>   <dbl>
##  1 Filly B~ Featur~ Drama      80 R       Indom~    2013       4      19    2013
##  2 The Dish Featur~ Drama     101 PG-13   Warne~    2001       3      14    2001
##  3 Waiting~ Featur~ Come~      84 R       Sony ~    1996       8      21    2001
##  4 The Age~ Featur~ Drama     139 PG      Colum~    1993      10       1    2001
##  5 Malevol~ Featur~ Horr~      90 R       Ancho~    2004       9      10    2005
##  6 Old Par~ Docume~ Docu~      78 Unrated Shcal~    2009       1      15    2010
##  7 Lady Ja~ Featur~ Drama     142 PG-13   Param~    1986       1       1    2003
##  8 Mad Dog~ Featur~ Drama      93 R       MGM/U~    1996      11       8    2004
##  9 Beauty ~ Docume~ Docu~      88 Unrated Indep~    2012       9       7    2013
## 10 The Sno~ Featur~ Drama     119 Unrated IFC F~    2012       3       2    2012
## # ... with 641 more rows, 22 more variables: 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>, and ...
```

Let’s understand the data

We have 651 observation and 32 variables in the given data set. The response variable “audience_score” is numerical, and the predictor variable are mixed with numerical & categorical variables. There are 6 variables which which has only two class/levels “Yes & No” which can be comverted into numerical(will be done at later stage)

```
str(movies)
```

```
## tibble [651 x 32] (S3: tbl_df/tbl/data.frame)
##  $ title           : chr [1:651] "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 [1:651] 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 [1:651] 2013 2001 1996 1993 2004 ...
##  $ thtr_rel_month  : num [1:651] 4 3 8 10 9 1 1 11 9 3 ...
##  $ thtr_rel_day    : num [1:651] 19 14 21 1 10 15 1 8 7 2 ...
##  $ dvd_rel_year    : num [1:651] 2013 2001 2001 2001 2005 ...
##  $ dvd_rel_month   : num [1:651] 7 8 8 11 4 4 2 3 1 8 ...
##  $ dvd_rel_day     : num [1:651] 30 28 21 6 19 20 18 2 21 14 ...
##  $ imdb_rating     : num [1:651] 5.5 7.3 7.6 7.2 5.1 7.8 7.2 5.5 7.5 6.6 ...
##  $ imdb_num_votes  : int [1:651] 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 [1:651] 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 [1:651] 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 [1:651] "Michael D. Olmos" "Rob Sitch" "Christopher Guest" "Martin Scorsese" ...
##  $ actor1          : chr [1:651] "Gina Rodriguez" "Sam Neill" "Christopher Guest" "Daniel Day-Lewis" ...
##  $ actor2          : chr [1:651] "Jenni Rivera" "Kevin Harrington" "Catherine O'Hara" "Michelle Pfeiffer" ...
##  $ actor3          : chr [1:651] "Lou Diamond Phillips" "Patrick Warburton" "Parker Posey" "Winona Ryder" ...
##  $ actor4          : chr [1:651] "Emilio Rivera" "Tom Long" "Eugene Levy" "Richard E. Grant" ...
##  $ actor5          : chr [1:651] "Joseph Julian Soria" "Genevieve Mooy" "Bob Balaban" "Alec McCowen" ...
##  $ imdb_url        : chr [1:651] "http://www.imdb.com/title/tt1869425/" "http://www.imdb.com/title/tt0205873/" "
##  $ rt_url          : chr [1:651] "//www.rottentomatoes.com/m/filly_brown_2012/" "//www.rottentomatoes.com/m/dish
```

```
summary(movies)
```

```
##     title                  title_type                 genre         runtime
##  Length:651         Documentary : 55   Drama             :305   Min.   : 39.0
##  Class :character   Feature Film:591   Comedy            : 87   1st Qu.: 92.0
##  Mode  :character   TV Movie    :  5   Action & Adventure: 65   Median :103.0
##                                        Mystery & Suspense: 59   Mean   :105.8
##                                        Documentary       : 52   3rd Qu.:115.8
##                                        Horror            : 23   Max.   :267.0
##                                        (Other)           : 60   NA's   :1
##   mpaa_rating                               studio    thtr_rel_year
##  G      : 19   Paramount Pictures            : 37   Min.   :1970
##  NC-17  :  2   Warner Bros. Pictures         : 30   1st Qu.:1990
##  PG     :118   Sony Pictures Home Entertainment: 27   Median :2000
##  PG-13  :133   Universal Pictures              : 23   Mean   :1998
##  R      :329   Warner Home Video               : 19   3rd Qu.:2007
##  Unrated: 50   (Other)                         :507   Max.   :2014
##                NA's                            :  8
##  thtr_rel_month   thtr_rel_day    dvd_rel_year  dvd_rel_month
##  Min.   : 1.00   Min.   : 1.00   Min.   :1991   Min.   : 1.000
##  1st Qu.: 4.00   1st Qu.: 7.00   1st Qu.:2001   1st Qu.: 3.000
##  Median : 7.00   Median :15.00   Median :2004   Median : 6.000
##  Mean   : 6.74   Mean   :14.42   Mean   :2004   Mean   : 6.333
##  3rd Qu.:10.00   3rd Qu.:21.00   3rd Qu.:2008   3rd Qu.: 9.000
##  Max.   :12.00   Max.   :31.00   Max.   :2015   Max.   :12.000
##                                  NA's   :8      NA's   :8
##   dvd_rel_day     imdb_rating    imdb_num_votes           critics_rating
##  Min.   : 1.00   Min.   :1.900   Min.   :   180   Certified Fresh:135
##  1st Qu.: 7.00   1st Qu.:5.900   1st Qu.:  4546   Fresh          :209
##  Median :15.00   Median :6.600   Median : 15116   Rotten         :307
##  Mean   :15.01   Mean   :6.493   Mean   : 57533
##  3rd Qu.:23.00   3rd Qu.:7.300   3rd Qu.: 58301
##  Max.   :31.00   Max.   :9.000   Max.   :893008
##  NA's   :8
##  critics_score    audience_rating audience_score  best_pic_nom best_pic_win
##  Min.   :  1.00   Spilled:275     Min.   :11.00   no :629      no :644
##  1st Qu.: 33.00   Upright:376     1st Qu.:46.00   yes: 22      yes:  7
##  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_actor_win best_actress_win best_dir_win top200_box   director
##  no :558        no :579          no :608      no :636    Length:651
##  yes: 93        yes: 72          yes: 43      yes: 15    Class :character
##                                                          Mode  :character
##
##
##
##
##     actor1             actor2             actor3             actor4
##  Length:651         Length:651         Length:651         Length:651
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     actor5             imdb_url            rt_url
##  Length:651         Length:651         Length:651
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

Exploratory Data Analysis

Check for the duplicate row/observations. Will be removed if there are any.

```
sum(duplicated(movies))
```

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

```
# Deleting the duplicate entry:
movies <- movies[!duplicated(movies), ]
```

Missing values

```
# Used 'colSums()' function to aggregate NA in each column
colSums(sapply(movies, is.na))
```

```
##            title       title_type            genre          runtime
##                0                0                0                1
##      mpaa_rating           studio    thtr_rel_year   thtr_rel_month
##                0                8                0                0
##     thtr_rel_day     dvd_rel_year    dvd_rel_month      dvd_rel_day
##                0                8                8                8
##      imdb_rating   imdb_num_votes   critics_rating    critics_score
##                0                0                0                0
##  audience_rating   audience_score     best_pic_nom     best_pic_win
##                0                0                0                0
##   best_actor_win best_actress_win     best_dir_win      top200_box
##                0                0                0                0
##         director           actor1           actor2           actor3
##                2                2                7                9
##           actor4           actor5         imdb_url           rt_url
##               13               15                0                0
```

Dropping observations with missing values

With less than 2% of the total observation, few details are missing. As the size is less, we are dropping the observations/row with missing values.

```
movies <- movies[!is.na(movies$runtime), ]
movies <- movies[!is.na(movies$dvd_rel_year), ]
movies <- movies[!is.na(movies$director), ]
movies <- movies[!is.na(movies$studio), ]
movies <- movies[!is.na(movies$actor1), ]
dim(movies)
```

```
## [1] 629  32
```

Dropping unwanted/irrelevant variables

1. Variables such as 'imdb_url' or 'rt_url' are excluded as they are not relevant to the purpose of identifying the popularity of the movie or imdb rating.
2. Other irrelevant variables such as title, actors, dvd release dates and studio are excluded from the model. 3.Theater release month was included assuming that movies released at certain times of the year may be more popular than others.

```
movies <- subset(movies, select = -c(imdb_url))
movies <- subset(movies, select = -c(rt_url))
#movies <- subset(movies, select = -c(title))
#movies <- subset(movies, select = -c(actor1))
#movies <- subset(movies, select = -c(actor2))
movies <- subset(movies, select = -c(actor3))
movies <- subset(movies, select = -c(actor4))
movies <- subset(movies, select = -c(actor5))
movies <- subset(movies, select = -c(dvd_rel_year))
movies <- subset(movies, select = -c(dvd_rel_month))
movies <- subset(movies, select = -c(dvd_rel_day))
#movies <- subset(movies, select = -c(studio))

dim(movies)
```

```
## [1] 629  24
```

Converting binary variables into numeric:

```
# Converting binary variable yes or no into numerical for calculating correlation between them.
movies$best_dir_win<-ifelse(movies$best_dir_win=="yes",1,0)
movies$best_actor_win<-ifelse(movies$best_actor_win=="yes",1,0)
movies$best_actress_win<-ifelse(movies$best_actress_win=="yes",1,0)
movies$best_pic_win<-ifelse(movies$best_pic_win=="yes",1,0)
movies$best_pic_win<-ifelse(movies$best_pic_nom=="yes",1,0)
```

Q1) Oscar win actor/actress or director influence the overall performance or rating of the movie or not? OR Is an Oscar winning actor or actress in the cast associated with a better rating (in the context of multiple regression)?

```r
# Model_1 created with Response variable as 'imdb_rating' and rest all the variables as predictor variable.

model <- lm(imdb_rating ~ genre + runtime + best_actor_win + best_actress_win
            + best_dir_win + mpaa_rating + studio, data = movies)
summary(model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + best_actor_win +
##     best_actress_win + best_dir_win + mpaa_rating + studio, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5950 -0.3113  0.0000  0.3340  2.3399
##
## Coefficients: (1 not defined because of singularities)
##                                                  Estimate Std. Error t value
## (Intercept)                                      4.609398   0.460863  10.002
## genreAnimation                                   0.041448   0.400149   0.104
## genreArt House & International                   0.828995   0.382457   2.168
## genreComedy                                      0.093441   0.168738   0.554
## genreDocumentary                                 1.745969   0.299286   5.834
## genreDrama                                       0.704006   0.150991   4.663
## genreHorror                                      0.127655   0.285427   0.447
## genreMusical & Performing Arts                   1.489438   0.328824   4.530
## genreMystery & Suspense                          0.417172   0.194781   2.142
## genreOther                                       0.933080   0.303100   3.078
## genreScience Fiction & Fantasy                   0.222777   0.440429   0.506
## runtime                                          0.014167   0.002813   5.035
## best_actor_win                                  -0.073911   0.125519  -0.589
## best_actress_win                                -0.083160   0.133465  -0.623
## best_dir_win                                     0.308087   0.166847   1.847
## mpaa_ratingNC-17                                 0.510762   0.959189   0.532
## mpaa_ratingPG                                   -0.467315   0.301959  -1.548
## mpaa_ratingPG-13                                -0.749871   0.316808  -2.367
## mpaa_ratingR                                    -0.348463   0.310263  -1.123
## mpaa_ratingUnrated                              -0.284820   0.450405  -0.632
## studio20th Century Fox Film Corporat             1.113544   0.679899   1.638
## studio20th Century Fox Film Corporation          0.191461   0.464489   0.412
## studio7-57 Releasing                             0.191736   0.979895   0.196
## studio905 Corporation                           -0.606609   0.913098  -0.664
## studioA24                                       -0.364776   0.916551  -0.398
## studioA24 Films                                  0.110558   0.667895   0.166
## studioAll Girl Productions                      -0.695525   0.913371  -0.761
## studioAlliance Atlantis Communications           0.390057   0.911404   0.428
## studioAmerican International Pictures           -1.476090   0.912855  -1.617
## studioAnalysis                                   0.302558   0.913295   0.331
## studioAnchor Bay Entertainment                   0.099150   0.676687   0.147
## studioAnchor Bay Films                           0.381409   0.942582   0.405
## studioArenas Entertainment                      -0.222605   0.985801  -0.226
## studioArtisan Entertainment                      0.613123   0.913888   0.671
## studioAVCO Embassy Pictures                      0.421636   0.922232   0.457
## studioBankside Films                             0.796241   0.991112   0.803
## studioBlumhouse                                -0.396609   0.912170  -0.435
## studioBMG                                       0.962299   0.914002   1.053
## studioBrainstorm Media                          1.041415   0.983314   1.059
## studioBuena Vista                              -0.231711   0.420168  -0.551
## studioBuena Vista Distribution Compa            1.979981   0.681974   2.903
## studioBuena Vista Internationa                  0.919224   0.911208   1.009
## studioBuena Vista Pictures                      0.388779   0.382370   1.017
## studioCarnaby International                     0.953230   0.916001   1.041
## studioChloe Productions                        -2.867443   0.911627  -3.145
## studioCine-Source                              -0.633532   0.917034  -0.691
## studioCinema Libre Studio                      -0.028493   1.018035  -0.028
## studioCinema Seven Productions Ltd              0.741249   0.913681   0.811
## studioCinetic Media                             0.296951   0.992283   0.299
## studioCode Red                                  1.085593   1.031705   1.052
## studioColumbia Pictures                         0.350510   0.404746   0.866
## studioColumbia Tristar Pictures                 1.022397   0.914958   1.117
## studioConcorde/New Horizons Home Video         -2.164424   0.942678  -2.296
## studioCowboy Pictures                           0.238088   0.693513   0.343
## studioCriterion Collection                      0.776427   0.950657   0.817
## studioCrown International Pictures             -2.429546   1.004414  -2.419
## studioD&E Entertainment                         0.097784   0.991266   0.099
## studioDestination Films                        -2.654109   0.912600  -2.908
## studioDimension Films                           0.034391   0.916454   0.038
## studioDisney                                    1.266410   0.912817   1.387
## studioDiva                                      0.308618   0.998007   0.309
## studioDreamworks                               -0.594775   0.916645  -0.649
## studioDreamWorks Studios                        1.468132   0.918639   1.598
## studioE1 Entertainment                         -2.175252   0.986572  -2.205
## studioEcho Bridge Home Entertainment            0.232557   0.911627   0.255
## studioEmbassy                                   1.630364   0.914068   1.784
## studioEmpire Pictures                           0.146992   0.987236   0.149
## studioeRealBiz                                  0.345891   0.912600   0.379
## studioFabrication Films                         0.945058   0.913907   1.034
## studioFathom Events                             1.234570   0.920059   1.342
## studioFilm Movement                            -0.059715   0.991654  -0.060
## studioFilmDistrict                              1.156515   0.664388   1.741
## studioFireworks Pictures                       -0.279943   0.911297  -0.307
## studioFirst Look                               -2.143544   0.913647  -2.346
## studioFirst Run Entertainment                   0.156951   0.990424   0.158
## studioFirst Run Features                       -0.140015   0.548265  -0.255
## studioFocus Features                            0.409321   0.502575   0.814
## studioFox                                       0.571024   0.571228   1.000
## studioFox Atomic                               -1.012706   0.664929  -1.523
## studioFox Searchlight                           0.733846   0.513599   1.429
## studioFox Searchlight Pictures                  0.664291   0.581242   1.143
## studioFreestyle Releasing                      -1.858695   0.918513  -2.024
## studioGener8Xion Entertainment                 -0.288590   0.914216  -0.316
## studioGenius Productions                       -0.320257   0.913097  -0.351
## studioGood Machine                              0.825304   0.924264   0.893
## studioGramercy Pictures                         1.248391   0.911228   1.370
## studioGravitas                                 -0.196609   0.912170  -0.216
## studioGravitas Ventures                         0.911415   0.984885   0.925
## studioGreyCat Films                                   NA         NA      NA
## studioGrindhouse Entertainment                  0.406564   0.916043   0.444
## studioGroup 1                                   0.127120   1.005000   0.126
## studioHatchet Films                            -0.524942   0.911927  -0.576
## studioHBO Documentary                          -0.002216   0.991266  -0.002
## studioHBO Video                                 0.397792   0.425365   0.935
## studioHemdale                                   0.349224   0.911622   0.383
## studioHK Film Corporation                      -1.410776   0.912044  -1.547
## studioHollywood Pictures                       -0.577340   0.558369  -1.034
## studioHouston Museum of Natural Scie            0.676952   1.009632   0.670
## studioIcarus Films                              0.254451   0.991800   0.257
## studioIFC                                       0.113542   0.668711   0.170
## studioIFC Films                                 0.250062   0.350791   0.713
## studioIFC First Take                            1.105891   0.911345   1.213
## studioIFC Midnight                              0.386933   1.005930   0.385
## studioImage Entertainment                      -0.324708   0.568832  -0.571
## studioIndependent Pictures                      0.278533   0.687210   0.405
## studioIndomina Films                            0.090594   0.950657   0.095
## studioIndomina Media Inc.                      -0.598276   0.914842  -0.654
## studioInnovation Film Group                     0.496735   0.979107   0.507
## studioIsland                                    0.732557   0.911627   0.804
## studioKaga Bay                                  0.336113   0.947932   0.355
## studioLaurel Entertainment                     -0.633276   0.917108  -0.691
## studioLevitt-Pickman Film Corporation          -0.356516   0.956076  -0.373
## studioLiberation Entertainment                 -0.058882   0.990860  -0.059
## studioLions Gate Films                          0.373826   0.412695   0.906
## studioLions Gate Films Inc.                     0.350576   0.942176   0.372
## studioLions Gate Releasing                      0.454909   0.680082   0.669
## studioLionsgate                                -0.686770   0.503023  -1.365
## studioLionsGate Entertainment                  -0.618471   0.697182  -0.887
## studioLionsgate Films                          -0.732952   0.558090  -1.313
## studioLionsgate Releasing                       0.417558   0.912305   0.458
## studioLive Home Video                           0.222302   0.917108   0.242
## studioLorimar Home Video                        0.851724   0.915347   0.930
## studioMadman Entertainment                      0.189761   0.950888   0.200
## studioMagic Lamp Releasing                     -0.467443   0.911627  -0.513
## studioMagnet Releasing                          0.063969   0.918081   0.070
## studioMagnet/Magnolia Pictures                  1.405891   0.911345   1.543
## studioMagnolia Pictures                         0.171128   0.394802   0.433
## studioMCA Universal Home Video                  0.284396   0.350014   0.813
## studioMedia Home Entertainment                  1.261058   0.917822   1.374
## studioMetro-Goldwyn-Mayer Pictures              1.307299   0.913293   1.431
## studioMGM                                      -0.308710   0.321436  -0.960
## studioMGM Home Entertainment                    0.158904   0.372380   0.427
## studioMGM/UA                                    0.825471   0.916738   0.900
## studioMGM/United Artists                       -0.708531   0.919692  -0.770
## studioMillenium Entertainment                  -0.694943   0.911194  -0.763
## studioMiramax                                  -0.037526   0.409098  -0.092
## studioMiramax Films                             0.471841   0.313317   1.506
## studioMusic Box Films                           0.866569   0.691149   1.254
## studioNational General Pictures                 0.718391   0.911544   0.788
## studioNational Geographic                       0.077965   0.999256   0.078
## studioNational Geographic Entertainment         1.014763   0.717518   1.414
## studioNelson Entertainment                     -0.598785   0.685573  -0.873
## studioNew Concorde Home Entertainment           0.220576   0.943482   0.234
## studioNew Line Cinema                          -0.460462   0.362252  -1.271
## studioNew Line Home Entertainment              -0.211673   0.408215  -0.519
## studioNew World Pictures                        1.246975   0.913306   1.365
## studioNew World Video                          -0.509943   0.911404  -0.560
## studioNew Yorker Films                          1.165584   0.574335   2.029
## studioNewmarket Film Group                      0.874051   0.917991   0.952
## studioNewmarket Films                           2.221058   0.916599   2.423
## studioNordisk Film Biograf Distribution         0.599748   0.982890   0.610
## studioNordisk Film Biografdistributi            1.060058   0.912761   1.161
## studioOctober Films                             0.253390   0.919932   0.275
## studioOpen Road Films                          -0.872710   0.913971  -0.955
## studioOrion                                   0.361466   0.915035   0.395
## studioOrion Home Video                       -0.205116   0.460844  -0.445
## studioOrion Pictures                          0.688217   0.918066   0.750
## studioOrion Pictures Corporation              0.266396   0.502604   0.530
## studioOutsider Films                          1.117725   0.918509   1.217
## studioOverture Films                          0.571920   0.667392   0.857
## studioParamount                              -1.063351   0.458127  -2.321
## studioParamount Classics                     -0.993573   0.951310  -1.044
## studioParamount Home Video                   -0.031241   0.354892  -0.088
## studioParamount Pictures                      0.385164   0.267404   1.440
## studioParamount Studios                       0.240657   0.499911   0.481
## studioParamount Vantage                       0.205057   0.911194   0.225
## studioPicturehouse                            0.290632   0.914254   0.318
## studioPolyGram Video                         -0.220375   0.922537  -0.239
## studioRelativity Media                        0.163558   0.916480   0.178
## studioRepublic Pictures Home Video            1.092224   0.666208   1.639
## studioRoadside Attraction                    -0.286678   0.958202  -0.299
## studioRoadside Attractions                    0.818675   0.666925   1.228
## studioSag Harbor-Basement Pictures            0.839451   0.992839   0.846
## studioSaguenay Films                          0.468618   0.991953   0.472
## studioSamuel Goldwyn Company                  1.275058   0.911927   1.398
## studioSamuel Goldwyn Films                    0.818047   0.664553   1.231
## studioScreen Gems                             0.295717   0.556704   0.531
## studioSeventh Art Productions                -0.283184   1.023271  -0.277
## studioShcalo Media Group                      0.624451   0.994163   0.628
## studioSnagFilms                              -2.053050   1.085927  -1.891
## studioSony Entertainment                      0.881410   0.912972   0.965
## studioSony Pictures                           0.228243   0.364275   0.627
## studioSony Pictures Classics                  0.765827   0.362459   2.113
## studioSony Pictures Entertainment             0.405683   0.503476   0.806
## studioSony Pictures Home Entertainment       -0.072826   0.285928  -0.255
## studioSony Pictures/Columbia                  0.416198   0.914094   0.455
## studioSony Pictures/Screen Gems               1.253881   0.664436   1.887
## studioStrand Releasing                        0.935438   0.770139   1.215
## studioSummit Entertainment                    0.717224   0.560896   1.279
## studioTango Entertainment                    -0.063276   0.913894  -0.069
## studioThe Film Arcade                         0.308302   0.921191   0.335
## studioThe Jerry Gross Organization            1.222242   0.942076   1.297
## studioThe Shooting Gallery                    1.190225   0.917158   1.298
## studioThe Weinstein Co.                       0.544770   0.568905   0.958
## studioThe Weinstein Company                   0.153553   0.419557   0.366
## studioThinkFilm                               0.105693   0.572901   0.184
## studioTouchstone Home Entertainment           0.825321   0.918220   0.899
## studioTouchstone Pictures                     0.292427   0.507525   0.576
## studioTribeca Films                           0.024748   0.986572   0.025
## studioTrimark                                 0.610083   0.667228   0.914
## studioTrimark Pictures                        0.992299   0.913218   1.087
## studioTriStar                                 1.280635   0.916997   1.397
## studioTriStar Pictures                        0.747711   0.664598   1.125
## studioTwentieth Century Fox Home Entertainment  0.267448   0.328087   0.815
## studioUnited Artists                          1.124433   0.466922   2.408
## studioUniversal                               0.272489   0.926959   0.294
## studioUniversal Pictures                      0.741480   0.290612   2.551
## studioUniversal Studios                       0.264250   0.666860   0.396
## studioUrban Vision Entertainment              2.079897   0.914915   2.273
## studioUSA Films                               0.310212   0.667534   0.465
## studioVestron Video                           0.682469   0.921259   0.741
## studioVirgin Vision                          -0.080776   0.911208  -0.089
## studioWalt Disney Home Entertainment          1.663658   1.043672   1.594
## studioWalt Disney Pictures                      1.112845   0.556673   1.999
## studioWalt Disney Productions                  -0.004545   0.677495  -0.007
## studioWarner Bros Pictures                      0.480799   0.920058   0.523
## studioWarner Bros.                              0.087156   0.425013   0.205
## studioWarner Bros. Pictures                     0.516751   0.277200   1.864
## studioWARNER BROTHERS PICTURES                 -0.052721   0.377204  -0.140
## studioWarner Home Video                         0.113523   0.305133   0.372
## studioWarner Independent                        0.419224   0.911208   0.460
## studioWarner Independent Pictures               0.515508   0.560135   0.920
## studioWarners Bros. Pictures                    0.303959   0.926798   0.328
## studioWeinstein Company                        -1.103312   0.672259  -1.641
## studioWinstar                                   1.060891   0.911818   1.163
## studioYari Film Group Releasing                 0.649799   0.913217   0.712
## studioZeitgeist Films                           0.367785   0.993248   0.370
##                                                Pr(>|t|)    
## (Intercept)                                     < 2e-16 ***
## genreAnimation                                  0.91755    
## genreArt House & International                  0.03078 *  
## genreComedy                                     0.58005    
## genreDocumentary                               1.11e-08 ***
## genreDrama                                     4.25e-06 ***
## genreHorror                                     0.65494    
## genreMusical & Performing Arts                 7.80e-06 ***
## genreMystery & Suspense                         0.03281 *  
## genreOther                                      0.00222 ** 
## genreScience Fiction & Fantasy                  0.61326    
## runtime                                        7.21e-07 ***
## best_actor_win                                  0.55630    
## best_actress_win                                0.53358    
## best_dir_win                                    0.06555 .  
## mpaa_ratingNC-17                                0.59468    
## mpaa_ratingPG                                   0.12250    
## mpaa_ratingPG-13                                0.01841 *  
## mpaa_ratingR                                    0.26205    
## mpaa_ratingUnrated                              0.52751    
## studio20th Century Fox Film Corporat            0.10224    
## studio20th Century Fox Film Corporation         0.68041    
## studio7-57 Releasing                            0.84497    
## studio905 Corporation                           0.50685    
## studioA24                                       0.69085    
## studioA24 Films                                 0.86861    
## studioAll Girl Productions                      0.44681    
## studioAlliance Atlantis Communications          0.66890    
## studioAmerican International Pictures           0.10666    
## studioAnalysis                                  0.74060    
## studioAnchor Bay Entertainment                  0.88358    
## studioAnchor Bay Films                          0.68595    
## studioArenas Entertainment                      0.82146    
## studioArtisan Entertainment                     0.50267    
## studioAVCO Embassy Pictures                     0.64778    
## studioBankside Films                            0.42223    
## studioBlumhouse                                 0.66394    
## studioBMG                                       0.29304    
## studioBrainstorm Media                          0.29019    
## studioBuena Vista                               0.58162    
## studioBuena Vista Distribution Compa            0.00390 ** 
## studioBuena Vista Internationa                  0.31368    
## studioBuena Vista Pictures                      0.30988    
## studioCarnaby International                     0.29866    
## studioChloe Productions                         0.00178 **
## studioCine-Source                               0.49006
## studioCinema Libre Studio                       0.97769
## studioCinema Seven Productions Ltd              0.41769
## studioCinetic Media                             0.76490
## studioCode Red                                  0.29332
## studioColumbia Pictures                         0.38700
## studioColumbia Tristar Pictures                 0.26448
## studioConcorde/New Horizons Home Video          0.02219 *
## studioCowboy Pictures                           0.73155
## studioCriterion Collection                      0.41457
## studioCrown International Pictures              0.01601 *
## studioD&E Entertainment                         0.92147
## studioDestination Films                         0.00384 **
## studioDimension Films                           0.97008
## studioDisney                                    0.16610
## studioDiva                                      0.75730
## studioDreamworks                                0.51680
## studioDreamWorks Studios                        0.11079
## studioE1 Entertainment                          0.02803 *
## studioEcho Bridge Home Entertainment            0.79877
## studioEmbassy                                   0.07524 .
## studioEmpire Pictures                           0.88171
## studioeRealBiz                                  0.70488
## studioFabrication Films                         0.30172
## studioFathom Events                             0.18040
## studioFilm Movement                             0.95201
## studioFilmDistrict                              0.08250 .
## studioFireworks Pictures                        0.75886
## studioFirst Look                                0.01945 *
## studioFirst Run Entertainment                   0.87417
## studioFirst Run Features                        0.79856
## studioFocus Features                            0.41587
## studioFox                                       0.31808
## studioFox Atomic                                0.12854
## studioFox Searchlight                           0.15383
## studioFox Searchlight Pictures                  0.25377
## studioFreestyle Releasing                       0.04367 *
## studioGener8Xion Entertainment                  0.75242
## studioGenius Productions                        0.72597
## studioGood Machine                              0.37243
## studioGramercy Pictures                         0.17145
## studioGravitas                                  0.82946
## studioGravitas Ventures                         0.35531
## studioGreyCat Films                                  NA
## studioGrindhouse Entertainment                  0.65741
## studioGroup 1                                   0.89941
## studioHatchet Films                             0.56518
## studioHBO Documentary                           0.99822
## studioHBO Video                                 0.35026
## studioHemdale                                   0.70186
## studioHK Film Corporation                       0.12269
## studioHollywood Pictures                        0.30177
## studioHouston Museum of Natural Scie            0.50293
## studioIcarus Films                              0.79765
## studioIFC                                       0.86526
## studioIFC Films                                 0.47635
## studioIFC First Take                            0.22566
## studioIFC Midnight                              0.70070
## studioImage Entertainment                       0.56843  
## studioIndependent Pictures                      0.68547  
## studioIndomina Films                            0.92413  
## studioIndomina Media Inc.                       0.51351  
## studioInnovation Film Group                     0.61220  
## studioIsland                                    0.42212  
## studioKaga Bay                                  0.72309  
## studioLaurel Entertainment                      0.49027  
## studioLevitt-Pickman Film Corporation           0.70942  
## studioLiberation Entertainment                  0.95264  
## studioLions Gate Films                          0.36557  
## studioLions Gate Films Inc.                     0.71002  
## studioLions Gate Releasing                      0.50394  
## studioLionsgate                                 0.17293  
## studioLionsGate Entertainment                   0.37555  
## studioLionsgate Films                           0.18982  
## studioLionsgate Releasing                       0.64742  
## studioLive Home Video                           0.80860  
## studioLorimar Home Video                        0.35267  
## studioMadman Entertainment                      0.84192  
## studioMagic Lamp Releasing                      0.60840  
## studioMagnet Releasing                          0.94449  
## studioMagnet/Magnolia Pictures                  0.12370  
## studioMagnolia Pictures                         0.66492  
## studioMCA Universal Home Video                  0.41697  
## studioMedia Home Entertainment                  0.17022  
## studioMetro-Goldwyn-Mayer Pictures              0.15309  
## studioMGM                                       0.33743  
## studioMGM Home Entertainment                    0.66981  
## studioMGM/UA                                    0.36842  
## studioMGM/United Artists                        0.44151  
## studioMillenium Entertainment                   0.44611  
## studioMiramax                                   0.92696  
## studioMiramax Films                             0.13286  
## studioMusic Box Films                           0.21064  
## studioNational General Pictures                 0.43110  
## studioNational Geographic                       0.93785  
## studioNational Geographic Entertainment         0.15806  
## studioNelson Entertainment                      0.38296  
## studioNew Concorde Home Entertainment           0.81527  
## studioNew Line Cinema                           0.20442  
## studioNew Line Home Entertainment               0.60437  
## studioNew World Pictures                        0.17291  
## studioNew World Video                           0.57612  
## studioNew Yorker Films                          0.04307 *
## studioNewmarket Film Group                      0.34160  
## studioNewmarket Films                           0.01583 *
## studioNordisk Film Biograf Distribution         0.54208  
## studioNordisk Film Biografdistributi            0.24618  
## studioOctober Films                             0.78312  
## studioOpen Road Films                           0.34022  
## studioOrion                                     0.69303  
## studioOrion Home Video                          0.65649  
## studioOrion Pictures                            0.45391  
## studioOrion Pictures Corporation                0.59638  
## studioOutsider Films                            0.22436  
## studioOverture Films                            0.39198  
## studioParamount                                 0.02078 *
## studioParamount Classics                        0.29691  
## studioParamount Home Video                      0.92990
## studioParamount Pictures                        0.15054
## studioParamount Studios                         0.63049
## studioParamount Vantage                         0.82206
## studioPicturehouse                              0.75073
## studioPolyGram Video                            0.81132
## studioRelativity Media                          0.85845
## studioRepublic Pictures Home Video              0.10190
## studioRoadside Attraction                       0.76495
## studioRoadside Attractions                      0.22034
## studioSag Harbor-Basement Pictures              0.39833
## studioSaguenay Films                            0.63688
## studioSamuel Goldwyn Company                     0.16282
## studioSamuel Goldwyn Films                      0.21905
## studioScreen Gems                               0.59558
## studioSeventh Art Productions                   0.78212
## studioShcalo Media Group                        0.53028
## studioSnagFilms                                 0.05940 .
## studioSony Entertainment                        0.33491
## studioSony Pictures                             0.53130
## studioSony Pictures Classics                    0.03523 *
## studioSony Pictures Entertainment               0.42085
## studioSony Pictures Home Entertainment          0.79908
## studioSony Pictures/Columbia                    0.64913
## studioSony Pictures/Screen Gems                 0.05986 .
## studioStrand Releasing                          0.22522
## studioSummit Entertainment                      0.20173
## studioTango Entertainment                       0.94483
## studioThe Film Arcade                           0.73804
## studioThe Jerry Gross Organization              0.19524
## studioThe Shooting Gallery                      0.19512
## studioThe Weinstein Co.                         0.33885
## studioThe Weinstein Company                     0.71457
## studioThinkFilm                                 0.85372
## studioTouchstone Home Entertainment             0.36928
## studioTouchstone Pictures                       0.56481
## studioTribeca Films                             0.98000
## studioTrimark                                   0.36108
## studioTrimark Pictures                          0.27786
## studioTriStar                                   0.16332
## studioTriStar Pictures                          0.26123
## studioTwentieth Century Fox Home Entertainment  0.41545
## studioUnited Artists                            0.01648 *
## studioUniversal                                 0.76894
## studioUniversal Pictures                        0.01110 *
## studioUniversal Studios                         0.69212
## studioUrban Vision Entertainment                0.02353 *
## studioUSA Films                                 0.64239
## studioVestron Video                             0.45925
## studioVirgin Vision                             0.92941
## studioWalt Disney Home Entertainment            0.11171
## studioWalt Disney Pictures                      0.04627 *
## studioWalt Disney Productions                   0.99465
## studioWarner Bros Pictures                      0.60156
## studioWarner Bros.                              0.83762
## studioWarner Bros. Pictures                     0.06302 .
## studioWARNER BROTHERS PICTURES                  0.88891
## studioWarner Home Video                         0.71005
## studioWarner Independent                        0.64571
## studioWarner Independent Pictures              0.35795
## studioWarners Bros. Pictures                   0.74311
## studioWeinstein Company                        0.10154
## studioWinstar                                  0.24532
## studioYari Film Group Releasing                0.47716
## studioZeitgeist Films                          0.71136
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8783 on 403 degrees of freedom
## Multiple R-squared:  0.5713, Adjusted R-squared:  0.3319
## F-statistic: 2.387 on 225 and 403 DF,  p-value: 1.577e-14
```

```
backwardeliminated_model <- step(model, direction = "backward", trace = FALSE)
```

```
summary(backwardeliminated_model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + best_dir_win + mpaa_rating,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8937 -0.5107  0.0696  0.6004  1.9741
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.260280   0.322165  16.328  < 2e-16 ***
## genreAnimation                  -0.361786   0.369376  -0.979 0.327745
## genreArt House & International  0.861493   0.294191   2.928 0.003535 **
## genreComedy                     -0.088221   0.153856  -0.573 0.566582
## genreDocumentary                 1.633640   0.205151   7.963 8.22e-15 ***
## genreDrama                       0.601197   0.130448   4.609 4.93e-06 ***
## genreHorror                     -0.148775   0.231690  -0.642 0.521029
## genreMusical & Performing Arts  1.140630   0.289987   3.933 9.34e-05 ***
## genreMystery & Suspense          0.387143   0.170127   2.276 0.023215 *
## genreOther                       0.721957   0.263777   2.737 0.006381 **
## genreScience Fiction & Fantasy  0.026754   0.342359   0.078 0.937738
## runtime                          0.012761   0.002069   6.168 1.26e-09 ***
## best_dir_win                     0.399552   0.149242   2.677 0.007623 **
## mpaa_ratingNC-17                 0.079323   0.945352   0.084 0.933157
## mpaa_ratingPG                   -0.620447   0.259058  -2.395 0.016920 *
## mpaa_ratingPG-13                -0.888431   0.263785  -3.368 0.000805 ***
## mpaa_ratingR                    -0.563579   0.256248  -2.199 0.028227 *
## mpaa_ratingUnrated              -0.449302   0.296845  -1.514 0.130647
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9097 on 611 degrees of freedom
## Multiple R-squared:  0.3027, Adjusted R-squared:  0.2833
## F-statistic:  15.6 on 17 and 611 DF,  p-value: < 2.2e-16
```

```
anova(model, backwardeliminated_model)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: imdb_rating ~ genre + runtime + best_actor_win + best_actress_win +
##     best_dir_win + mpaa_rating + studio
## Model 2: imdb_rating ~ genre + runtime + best_dir_win + mpaa_rating
##   Res.Df    RSS   Df Sum of Sq      F Pr(>F)
## 1    403 310.89
## 2    611 505.61 -208   -194.72 1.2135 0.0516 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Conclusion: As per the above base model, along with best actor, best actress and best director we have considered other variables like genre, runtime, studio & mpa rating to create a model where “imdb_rating” was my response variable. The Rˆ2 and adjusted Rˆ2 is 0.57 & 0.32 respectively. Whereas, post eliminating few variables based on backward elimination function, model finalized with genre, runtime, best director & mpaa rating as their final variables. The Rˆ2 and adjusted Rˆ2 is 0.30 & 0.28 respectively. With backward elimination(lowest AIC), it clearly evident that the best actor or actress are not influencing the performance of the movie, it is the Oscar wining director.

Q2)Is critics' rating associated with audience's rating, in the context of regression model with other potential predictors.

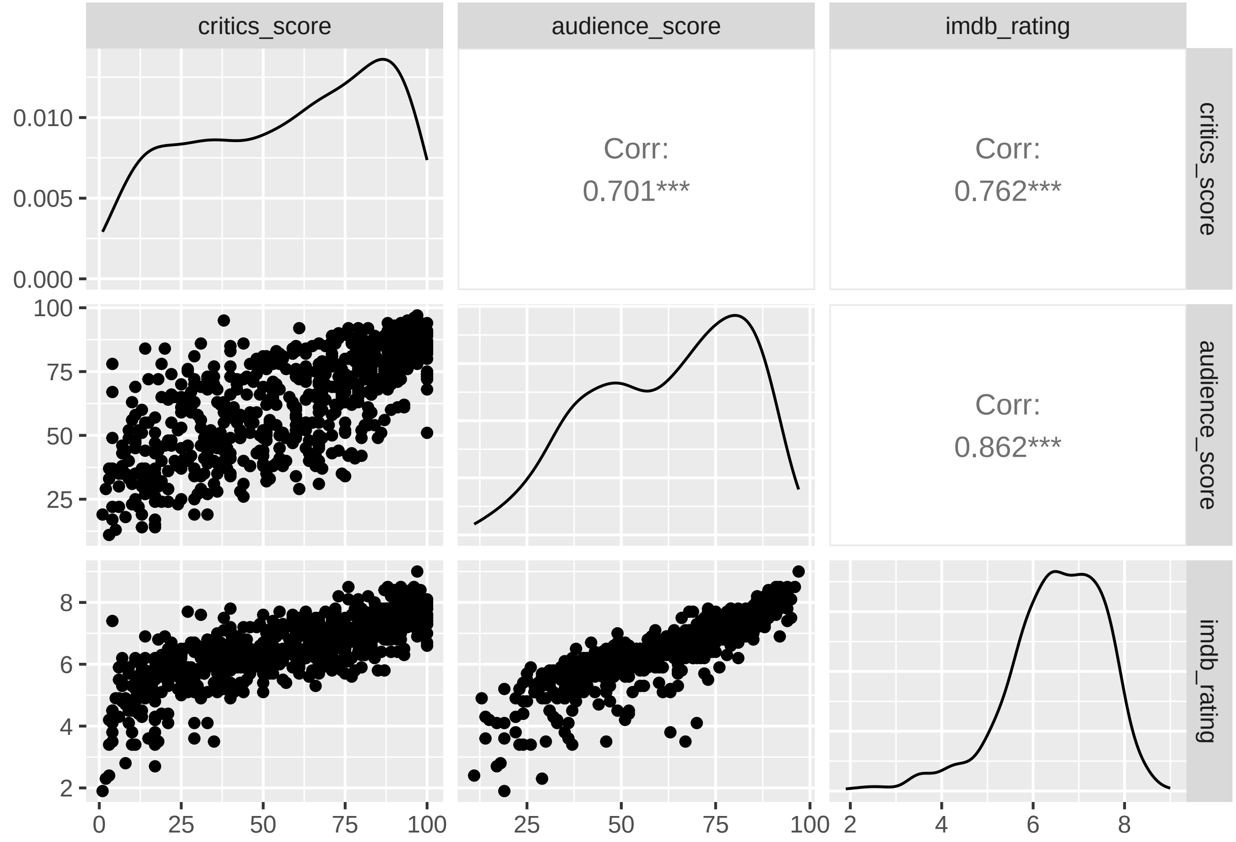
Correlation between audience_score and critics_score:

```
cor(movies$critics_score, movies$audience_score, use = 'everything', method = c('pearson'))
```

```
## [1] 0.7007955
```

Correlation between audience_score, critics_score and imdb_rating:

```
ggpairs(movies[, c('critics_score', 'audience_score', 'imdb_rating')])
```



Conclusion: 1.Correlation below is calculated based on Pearson's correlation. 2. With a correlation coefficient as high as 0.70, it is clear that critics and audience score most of the times agree with each other on the movie rating. 3. Also their is high correlation of 0.862 between audience_score and imdb_rating.

Q3) Create additional variable ‘Oscar’ with a class “yes or no” to identify whether an movie has won atleast one Oscar award or not and create a model with other variable along with ‘oscar’. Check if there is any significant difference after adding ‘Oscar’ in the model with other variable or not?

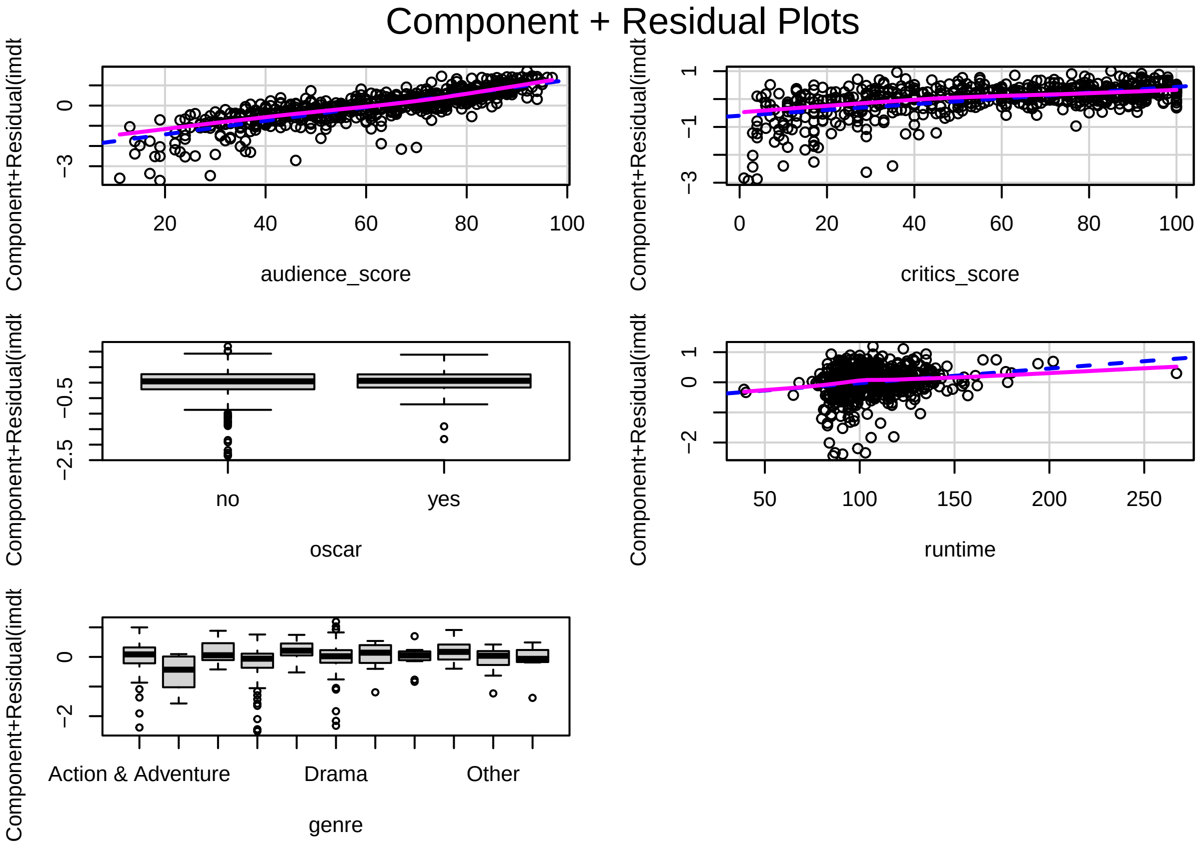
```
movies$oscar <- ifelse(movies$best_actor_win == 1 | movies$best_actress_win == 1 | movies$best_dir_win == 1, "yes",

model_less_oscar <- lm(imdb_rating ~ audience_score + critics_score, data=movies)
summary(model_less_oscar)
```

```
##
## Call:
## lm(formula = imdb_rating ~ audience_score + critics_score, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52421 -0.19533  0.03551  0.30294  1.21794
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.6731276  0.0640402   57.36   <2e-16 ***
## audience_score 0.0343878  0.0013660   25.18   <2e-16 ***
## critics_score  0.0117757  0.0009708   12.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4915 on 626 degrees of freedom
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7908
## F-statistic:  1188 on 2 and 626 DF,  p-value: < 2.2e-16
```

```
model_with_oscar <- lm(imdb_rating ~ audience_score + critics_score + oscar + runtime + genre, data=movies)
summary(model_with_oscar)
```

```
##
## Call:
## lm(formula = imdb_rating ~ audience_score + critics_score + oscar +
##     runtime + genre, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34358 -0.19874  0.04208  0.26617  1.18688
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.2300444  0.1298416  24.877  < 2e-16 ***
## audience_score                 0.0336188  0.0013439  25.015  < 2e-16 ***
## critics_score                  0.0103523  0.0009499  10.899  < 2e-16 ***
## oscaryes                       0.0634576  0.0452808   1.401  0.16159
## runtime                        0.0048592  0.0010763   4.515  7.6e-06 ***
## genreAnimation                -0.5040025  0.1762376  -2.860  0.00438 **
## genreArt House & International  0.2252544  0.1478337   1.524  0.12810
## genreComedy                   -0.1543898  0.0780216  -1.979  0.04828 *
## genreDocumentary               0.2687843  0.0977450   2.750  0.00614 **
## genreDrama                     0.0431389  0.0669821   0.644  0.51979
## genreHorror                    0.0895840  0.1166377   0.768  0.44275
## genreMusical & Performing Arts  0.0178641  0.1497888   0.119  0.90511
## genreMystery & Suspense         0.2346431  0.0863064   2.719  0.00674 **
## genreOther                     -0.0238874  0.1355736  -0.176  0.86020
## genreScience Fiction & Fantasy -0.0713931  0.1753835  -0.407  0.68410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4657 on 614 degrees of freedom
## Multiple R-squared:  0.8163, Adjusted R-squared:  0.8122
## F-statistic: 194.9 on 14 and 614 DF,  p-value: < 2.2e-16
```

```
library(car)
crPlots(model_with_oscar)
```



Conclusion: 1. A model with audience_score and critics_scoreas as a predictor accounts for 79.2 % (R-squared: 0.792, Adjusted R-squared: 0.7913) of the variation in imdb_rating. 2. Adding the categorical variable 'oscar' improves this to only 81.2 %(R-squared: 0.8168, Adjusted R-squared: 0.8127). 3. However, we were interested in quantifying the difference between imdb_rating of movies with a director, actor or actress who has won an oscar award compared to one without an oscar award. 4. The explanatory variables in our model (audience_score, critics_score, oscar, runtime and genre) are significant predictors for the response variable (imdb_rating).

Q4) Create additional variable 'movie_score' which is a combination of imdb_rating & audience_score to check if they add more weightage to the response variable 'imdb_rating'

```
movie_score = ((movies$imdb_rating*10)+movies$audience_score)/2
movies_1 <- movies %>% mutate(score = movie_score)
```

```
model_score_oscar <- lm(imdb_rating ~ movie_score + oscar, data=movies_1)
summary(model_score_oscar)
```

```
##
## Call:
## lm(formula = imdb_rating ~ movie_score + oscar, data = movies_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11445 -0.13960  0.02973  0.21180  0.85848
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.1768522  0.0639996  34.014  < 2e-16 ***
## movie_score 0.0674039  0.0009755  69.094  < 2e-16 ***
## oscaryes    0.1063455  0.0328629   3.236  0.00128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3648 on 626 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8847
## F-statistic:  2411 on 2 and 626 DF,  p-value: < 2.2e-16
```

```
summary(model_score_oscar)$adj.r.squared
```

```
## [1] 0.8847405
```

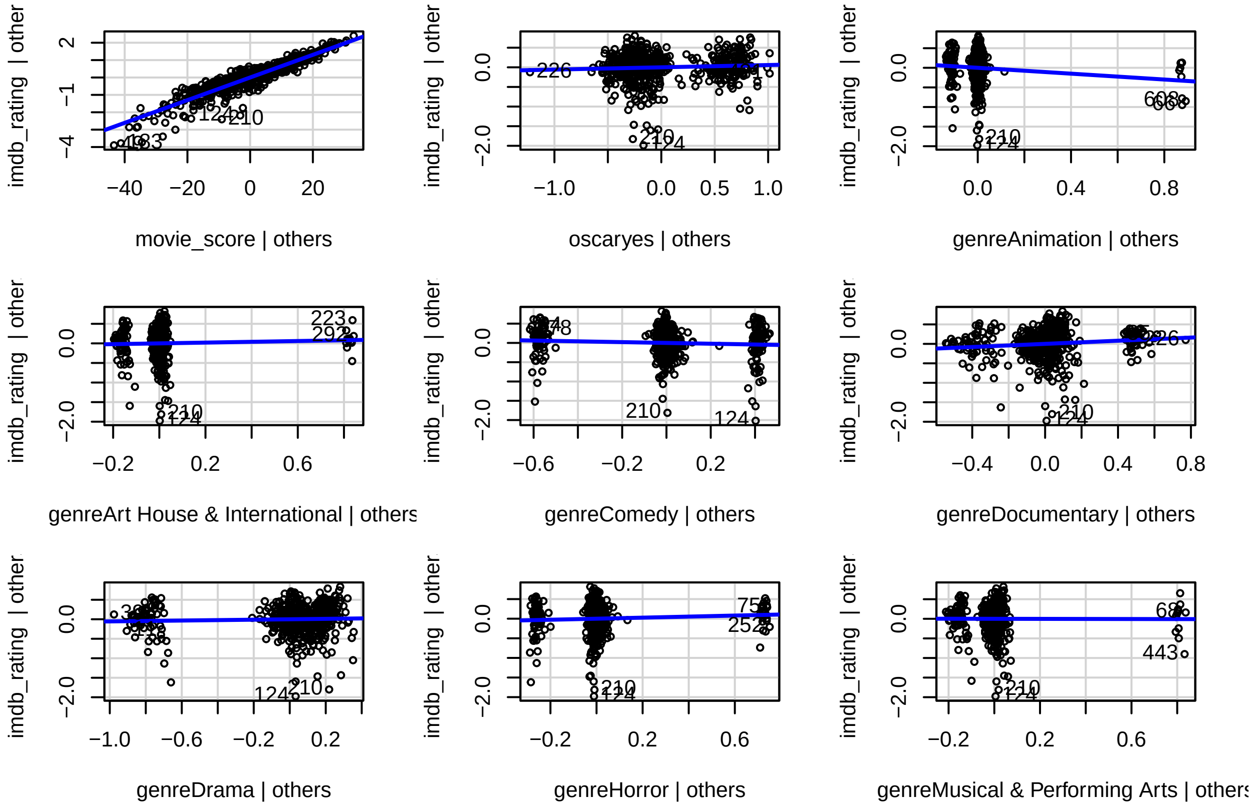
```
rev_model_score_oscar <- lm(imdb_rating ~ movie_score + oscar + genre + thtr_rel_month + runtime, data=movies_1)
summary(rev_model_score_oscar)
```

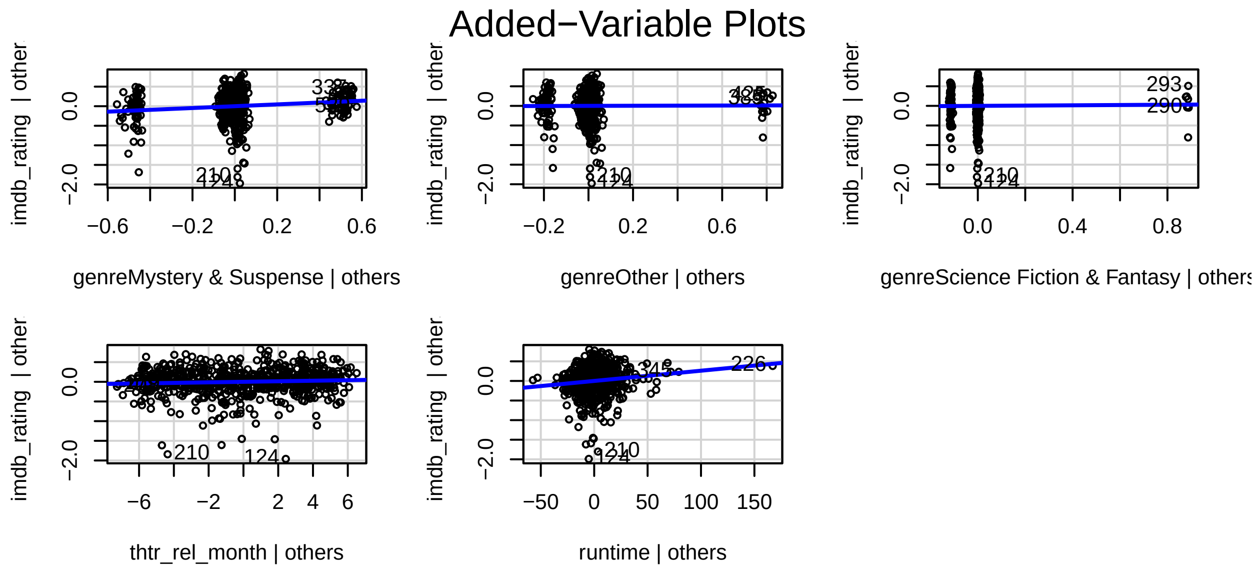
```
##
## Call:
## lm(formula = imdb_rating ~ movie_score + oscar + genre + thtr_rel_month +
##     runtime, data = movies_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97755 -0.14259  0.02592  0.20305  0.81722
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.9548389  0.1030771  18.965  < 2e-16 ***
## movie_score                     0.0652250  0.0010696  60.979  < 2e-16 ***
## oscaryes                        0.0581092  0.0338720   1.716 0.086748 .
## genreAnimation                 -0.3742273  0.1317779  -2.840 0.004663 **
## genreArt House & International  0.1000643  0.1107549   0.903 0.366628
## genreComedy                    -0.1029324  0.0584377  -1.761 0.078668 .
## genreDocumentary                0.2017705  0.0725924   2.780 0.005611 **
## genreDrama                      0.0546617  0.0498230   1.097 0.273019
## genreHorror                     0.1331592  0.0871002   1.529 0.126828
## genreMusical & Performing Arts -0.0086405  0.1119129  -0.077 0.938484
## genreMystery & Suspense         0.2321392  0.0643238   3.609 0.000333 ***
## genreOther                      0.0149965  0.1014183   0.148 0.882496
## genreScience Fiction & Fantasy  0.0363848  0.1310045   0.278 0.781308
## thtr_rel_month                  0.0068163  0.0040727   1.674 0.094705 .
## runtime                         0.0026134  0.0008287   3.154 0.001691 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3486 on 614 degrees of freedom
## Multiple R-squared:  0.8971, Adjusted R-squared:  0.8948
## F-statistic: 382.5 on 14 and 614 DF,  p-value: < 2.2e-16
```

```
summary(rev_model_score_oscar)$adj.r.squared
```

```
## [1] 0.8947846
```

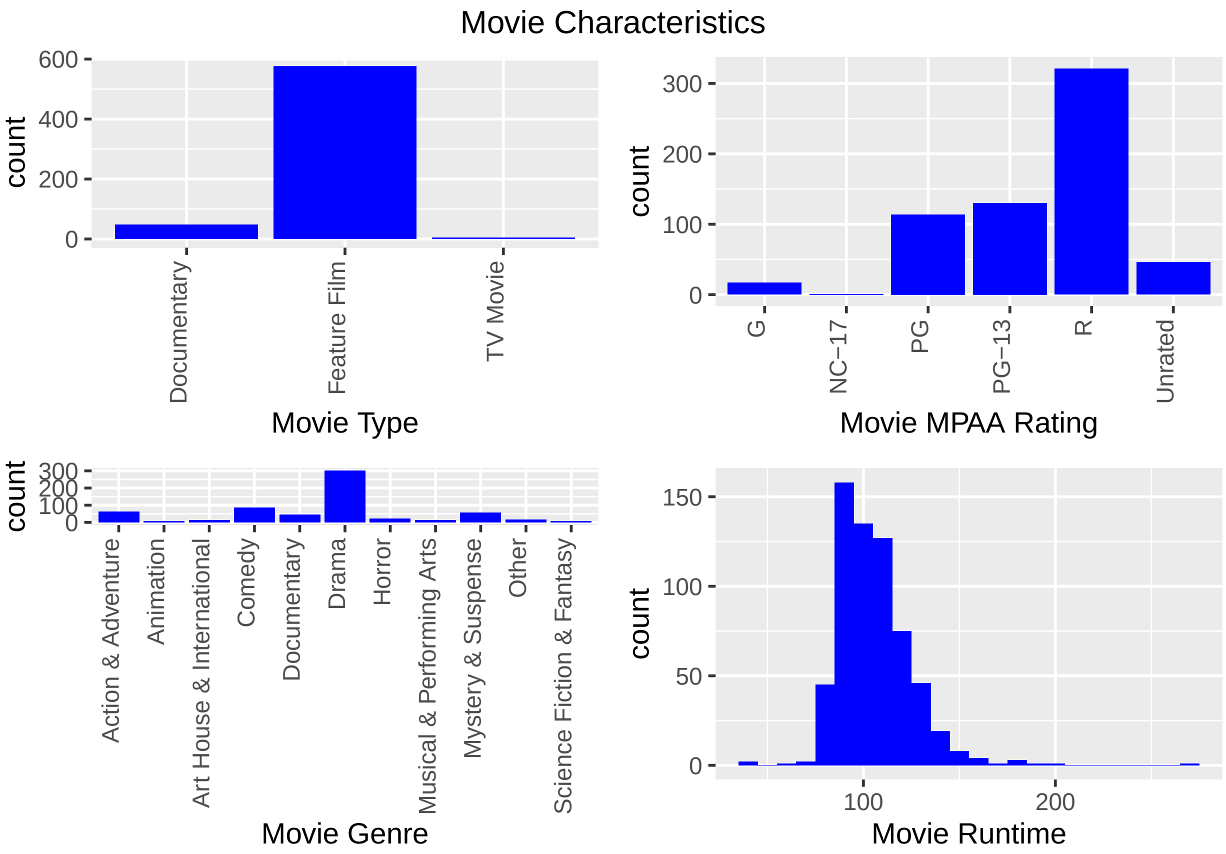
```
library(car)
avPlots(rev_model_score_oscar)
```




Added−Variable Plots
imdb_rating | other
0.0
−2.0
−0.6
−0.2
0.2
0.6
337
210
124
genreMystery & Suspense | others
imdb_rating | other
−0.2
0.2
0.6
125
389
210
124
genreOther | others
imdb_rating | other
0.0
0.4
0.8
293
290
210
124
genreScience Fiction & Fantasy | others
imdb_rating | other
−6
−2
2
4
6
210
124
thtr_rel_month | others
imdb_rating | other
−50
0
50
100
150
345
226
210
124
runtime | others

Q5) Insights about movies characteristics with reference to title_type, mpaa_rating or genre.

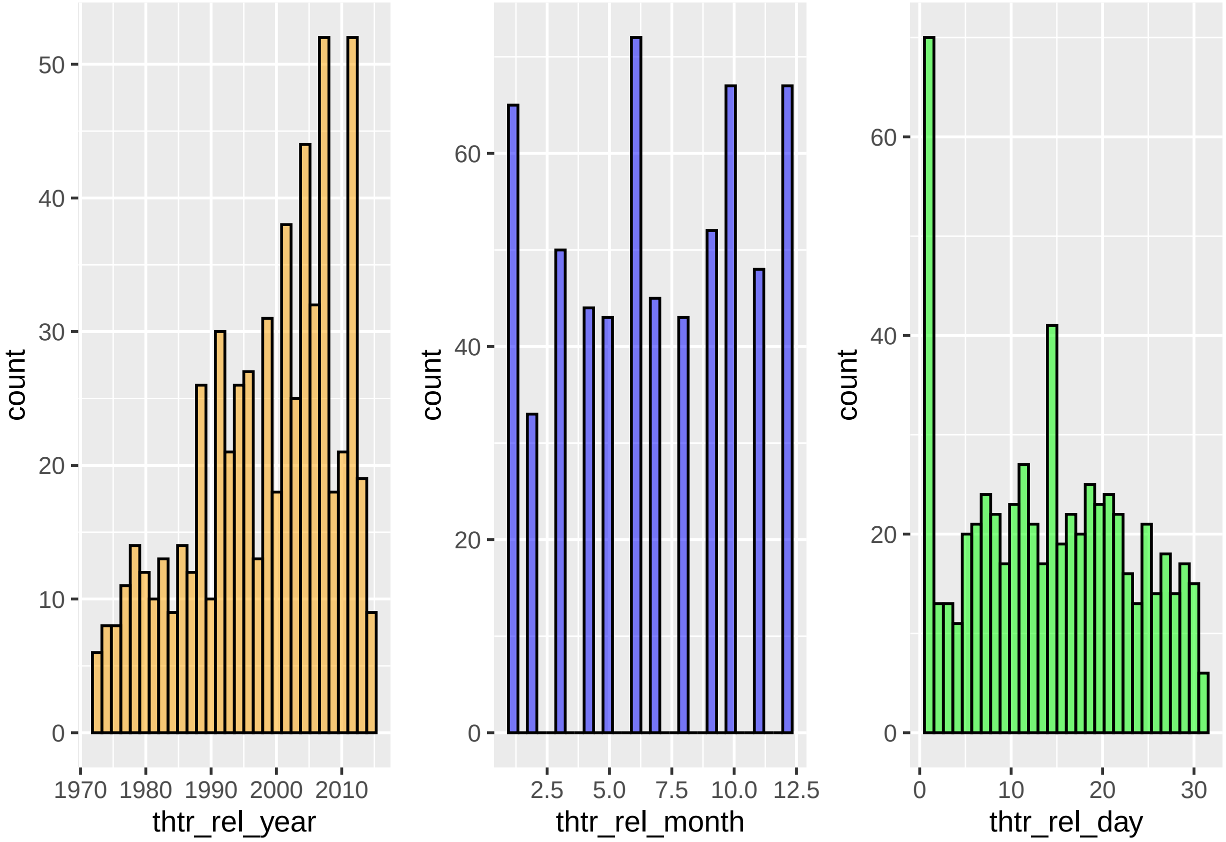
```
p1 <- ggplot(data=movies, aes(x=genre)) + geom_bar(fill = "blue") +
  xlab("Movie Genre") + theme(axis.text.x=element_text(angle=90,
                                                       hjust=1, vjust=0))

p2 <- ggplot(data=movies, aes(x=title_type)) + geom_bar(fill="blue") +
      xlab("Movie Type") +theme(axis.text.x=element_text(angle=90,
                                                         hjust=1, vjust=0))

p3 <- ggplot(data=movies, aes(x=mpaa_rating)) + geom_bar(fill="blue") +
      xlab("Movie MPAA Rating") +
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

p4 <- ggplot(data=movies, aes(x=runtime)) +geom_histogram(
  binwidth=10, fill="blue") + xlab("Movie Runtime")

grid.arrange(p2, p3, p1, p4, nrow=2, top="Movie Characteristics")
```



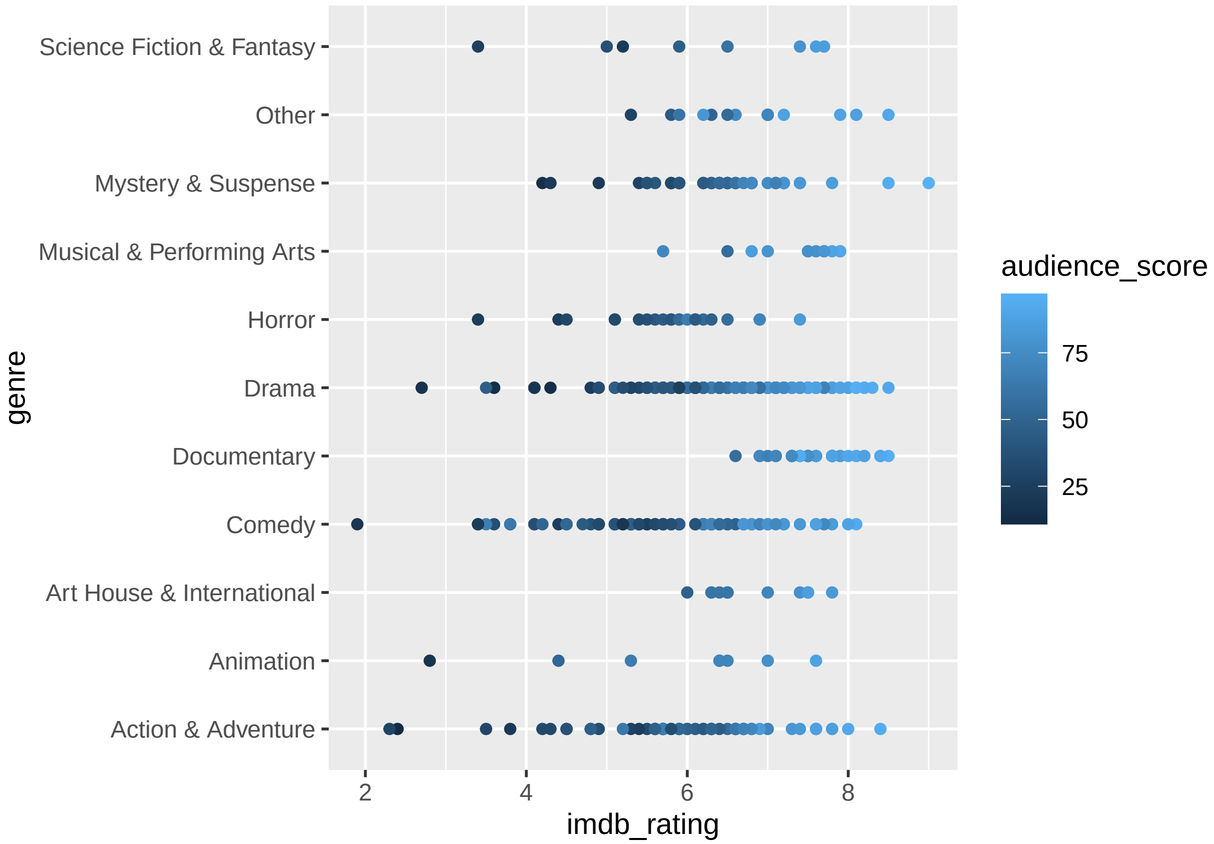
```
library(ggplot2)
g1 <- ggplot(data = movies, aes(x = thtr_rel_year)) + geom_histogram(colour = "black", fill = "orange", alpha = 0.5
g2 <- ggplot(data = movies, aes(x = thtr_rel_month)) + geom_histogram(colour = "black", fill = "blue", alpha = 0.5)
g3 <- ggplot(data = movies, aes(x = thtr_rel_day)) + geom_histogram(colour = "black", fill = "green", alpha = 0.5)

library(gridExtra)
```

```
grid.arrange(g1, g2, g3, nrow = 1, ncol = 3)
```

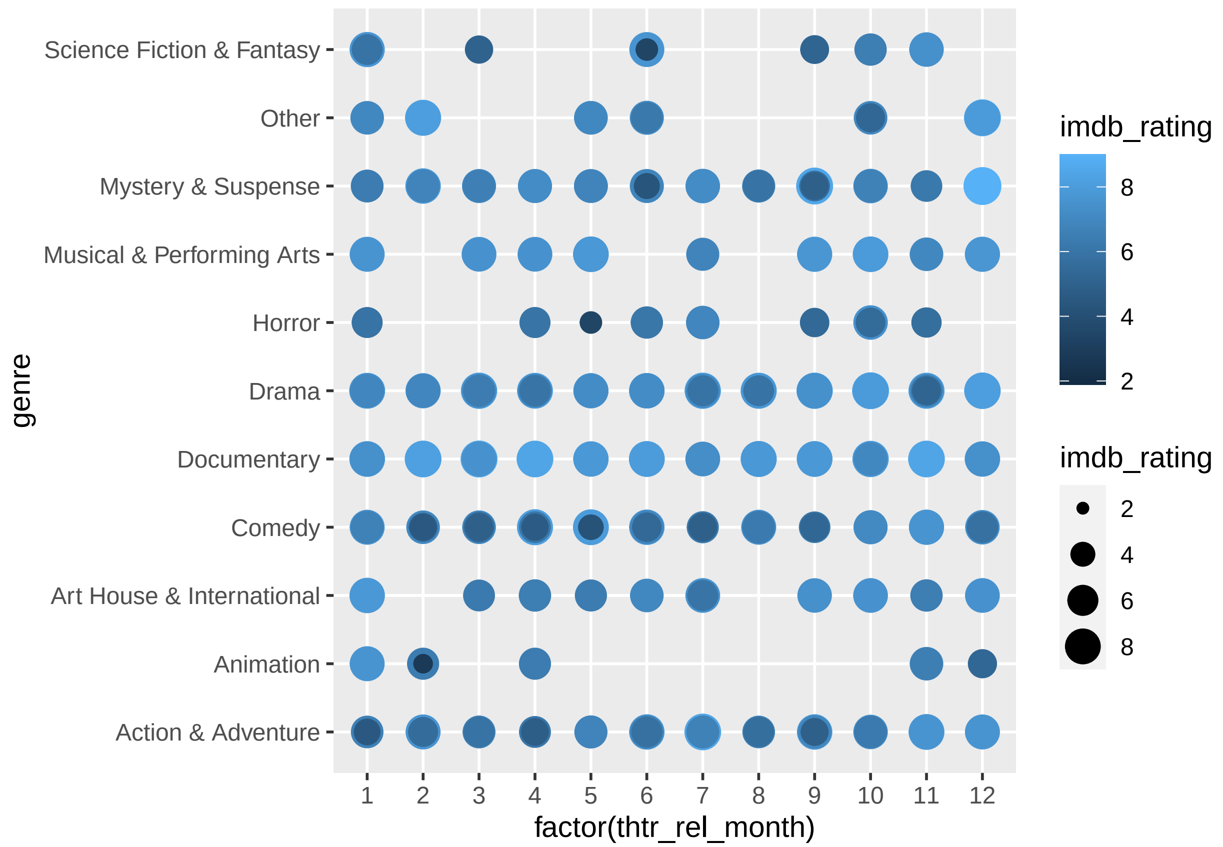
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
g4 <- ggplot(data = movies, aes(y = genre , x = imdb_rating, colour = audience_score)) + geom_point()
g4
```



```
g5 <- ggplot(data = movies, aes(y = genre , x = factor(thtr_rel_month), size = imdb_rating, colour = imdb_rating))
g5
```



Conclusion: 1. More movies are created and released under "Feature Film" title_type. 2. Major movies released belongs to mpaa_rating "R" & "PG-13 & PG" compared to others. 3. Also viewers like to watch Action & Adventure, Comedy,Drama or Mstery & Suspense types/genres of movies. 4. Movies are more released in the vacation season like in the month of Jan, Jun, Oct & Dec.

Q6) Prediction

```
predict_model = lm(imdb_rating ~ imdb_num_votes + genre + runtime + actor1 +  mpaa_rating , data=movies)
summary(predict_model)$r.squared
```

```
## [1] 0.9100267
```

```
summary(predict_model)$adj.r.squared
```

```
## [1] 0.60209
```

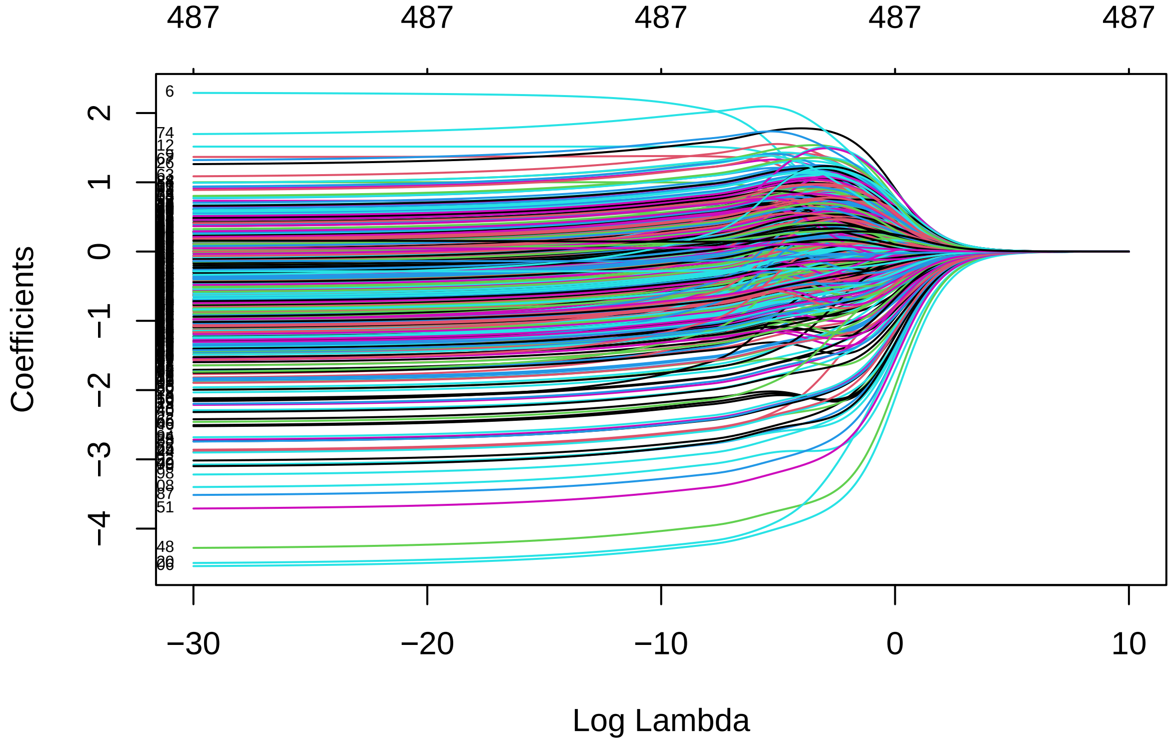
```
#summary(predict_model)
```

```
X1 = model.matrix(lm(imdb_rating ~ imdb_num_votes + genre + runtime + actor1 +  mpaa_rating, data = movies))
```

```
library(glmnet)
```

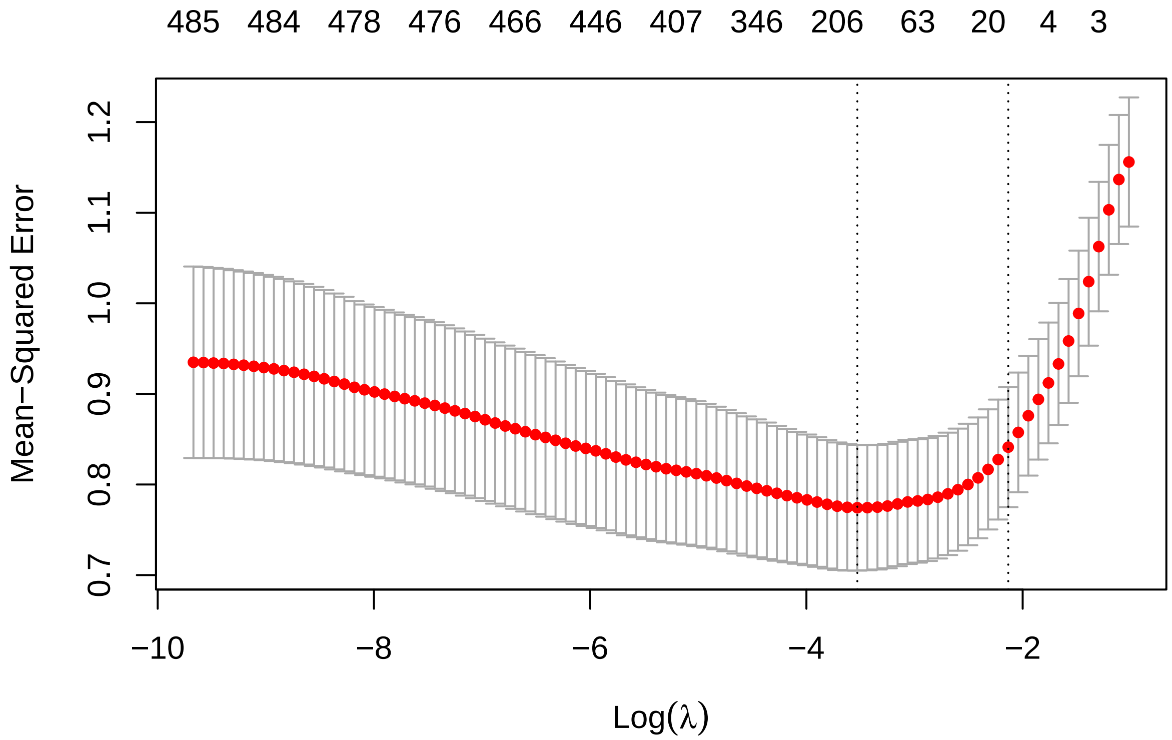
```
## Warning: package 'glmnet' was built under R version 4.2.2

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.2.2

## Loaded glmnet 4.1-6
```

```
ridge <- glmnet(X1, movies$imdb_rating, family = 'gaussian',
                lambda = exp(seq(-30,10,length.out=1000)), alpha = 0)

G1 = plot(ridge, xvar='lambda', label=TRUE)
```



The higher the lambda value , the more close it becomes to zero

```
library(glmnet)
CV1 = cv.glmnet(X1, movies$imdb_rating)
plot(CV1)
```



```
cv.glmnet(X1, movies$imdb_rating)$lambda.min
```

```
## [1] 0.03218914
```

```
cv.glmnet(X1, movies$imdb_rating)$lambda.1se
```

```
## [1] 0.1184039
```

```
rev_predict_model=lm(imdb_rating ~ X1 , data=movies)
summary(rev_predict_model)$r.squared
```

```
## [1] 0.9100267
```

```
summary(rev_predict_model)$adj.r.squared
```

```
## [1] 0.60209
```

```
#summary(predict_model)
```

```
#library(tidyverse)
fit <- lm(rev_predict_model, movies)

residual_standard_error <- summary(fit)$sigma
mean_of_response_variable <- mean(movies$imdb_rating)
error_percentage = (residual_standard_error / mean_of_response_variable) * 100
print(error_percentage)
```

```
## [1] 10.42617
```

```
#orgdf %>% filter(title=="Old Partner") %>%
#  select(c(genre,imdb_num_votes,actor1,mpaa_rating,runtime))

library(knitr)
data_1 <- data.frame(genre="Mystery & Suspense",  imdb_num_votes=6247, actor1="Gene Hackman", mpaa_rating = "R", ru
pred_1 <- predict(rev_predict_model, data_1, interval="predict")
```

```
## Warning: 'newdata' had 1 row but variables found have 629 rows
```

```
## Warning in predict.lm(rev_predict_model, data_1, interval = "predict"):
## prediction from a rank-deficient fit may be misleading
```

```
data_2 <- data.frame(genre="Documentary",  imdb_num_votes=333, actor1="Choi Won-kyun", mpaa_rating = "Unrated", run
pred_2 <- predict(rev_predict_model, data_2, interval="predict")
```

```
## Warning: 'newdata' had 1 row but variables found have 629 rows
```

```
## Warning in predict.lm(rev_predict_model, data_2, interval = "predict"):
## prediction from a rank-deficient fit may be misleading
```

```
# Show prediction results.
df <- data.frame(a=c("The Package  -  ", "Old Partner  -  "),
                 b=c(sprintf("%2.1f", pred_1[1]),
                     sprintf("%2.1f", pred_2[1])),
                 d=c(sprintf("%2.1f - %2.1f", pred_1[1,2], pred_1[1,3]),
                     sprintf("%2.1f - %2.1f", pred_2[1,2], pred_2[1,3])),
                 e=c("6.4", "7.8"))
kable(df, col.names=c("Movie Title", "Predicted Rating", "95% Prediction Interval", "Actual Rating"))
```

Movie Title	Predicted Rating	95% Prediction Interval	Actual Rating
The Package -	5.5	3.6 - 7.4	6.4
Old Partner -	5.5	3.6 - 7.4	7.8

```
#orgdf %>% filter(title=="Old Partner") %>%
#  select(c(genre,imdb_num_votes,actor1,mpaa_rating,runtime))

library(knitr)
data_1 <- data.frame(genre="Mystery & Suspense",  imdb_num_votes=6247, actor1="Gene Hackman", mpaa_rating = "R", ru
pred_1 <- predict(predict_model, data_1, interval="predict")
```

```
## Warning in predict.lm(predict_model, data_1, interval = "predict"): prediction
## from a rank-deficient fit may be misleading
```

```
data_2 <- data.frame(genre="Documentary",  imdb_num_votes=333, actor1="Choi Won-kyun", mpaa_rating = "Unrated", run
pred_2 <- predict(predict_model, data_2, interval="predict")
```

```
## Warning in predict.lm(predict_model, data_2, interval = "predict"): prediction
## from a rank-deficient fit may be misleading
```

```r
# Show prediction results.
df <- data.frame(a=c("The Package  -  ", "Old Partner  -  "),
                 b=c(sprintf("%2.1f", pred_1[1]),
                     sprintf("%2.1f", pred_2[1])),
                 d=c(sprintf("%2.1f - %2.1f", pred_1[1,2], pred_1[1,3]),
                     sprintf("%2.1f - %2.1f", pred_2[1,2], pred_2[1,3])),
                 e=c("6.4", "7.8"))
kable(df, col.names=c("Movie Title", "Predicted Rating", "95% Prediction Interval", "Actual Rating"))
```

Movie Title	Predicted Rating	95% Prediction Interval	Actual Rating
The Package -	6.7	5.1 - 8.2	6.4
Old Partner -	7.8	5.9 - 9.7	7.8