NBCUniversal - Box Office Forecasting

Chunzi Wang

March 20, 2018

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(xlsx)
```

```
movies <- read.xlsx2("C:/Users/Adimn/Desktop/NBCU-dataLaurel.xlsx",
                     sheetIndex=1,
                     colClasses = c("character","character","character","character","numeric","numeric"
```

```
dim(movies)
```

```
## [1] 8468   19
```

```
str(movies)
```

```
## 'data.frame':    8468 obs. of  19 variables:
##  $ imdbid          : Factor w/ 8468 levels "tt0008133","tt0010323",..: 2 9 13 26 74 80 99 145 217 247
##  $ title           : Factor w/ 8228 levels "'71","'night, Mother",..: 6361 2959 543 7683 3915 3682 25
##  $ plot            : Factor w/ 8163 levels "'4242' is the story of a young female immigrant, a story
##  $ rating          : Factor w/ 19 levels "","APPROVED",..: 18 7 7 5 5 5 9 18 3 5 ...
##  $ imdb_rating     : num  8.1 8 7.2 7.9 6.3 7.5 7.8 7.3 8 7.6 ...
##  $ metacritic      : num  NaN NaN NaN NaN NaN 95 67 72 95 82 ...
##  $ dvd_release     : Date, format: "1997-10-15" "2003-06-24" ...
##  $ production      : Factor w/ 1834 levels "","1-800-Love",..: 1362 1362 1362 396 1045 413 394 870 29
##  $ actors          : Factor w/ 8131 levels "","'Bad' Chad Broussard, Jere' Folley Chaisson, 'Crazy' '
##  $ imdb_votes      : num  42583 21154 17801 5705 132 ...
##  $ poster          : Factor w/ 7967 levels "","http://ia.media-imdb.com/images/M/MV5BM2RlNTFhZmEtZDR
##  $ director        : Factor w/ 5877 levels "","A. Sarkunam",..: 4677 121 2380 1702 2995 4642 2171 340
##  $ release_date    : Date, format: "1921-03-19" "1960-05-16" ...
##  $ runtime         : Factor w/ 194 levels "","1 min","10 min",..: 158 184 193 20 179 178 9 76 177 10
##  $ genre           : Factor w/ 540 levels "","Action","Action, Adventure",..: 453 426 424 202 202 264
##  $ awards          : Factor w/ 1144 levels "","1 nomination.",..: 2 777 26 14 493 213 904 2 1091 368
##  $ keywords        : Factor w/ 6166 levels "","\\N","007|terrorist-cell|intelligence-agency|computer-
##  $ Budget          : num  18000 88300 220000 6590 0 ...
##  $ Box.Office.Gross: num  0 0 46585 0 0 ...
```

```
#head(movies)
```

Remove 'plot' and 'poster' column because they are not needed here.

```
movies <- movies[,-c(3,11)]
summary(movies)
```

```
##        imdbid                          title           rating
##  tt0008133:   1   Untitled Fox/Marvel Film:   7   N/A      :3213
##  tt0010323:   1   Black Mass              :   3   R        :2237
##  tt0028212:   1   Cinderella              :   3   PG-13    :1344
##  tt0028358:   1   Sleepless               :   3   NOT RATED: 683
##  tt0042332:   1   The Night Before        :   3   PG       : 615
##  tt0046004:   1   The Switch              :   3   UNRATED  : 187
```

```
##  (Other)  :8462   (Other)                :8446   (Other)  : 189
##   imdb_rating       metacritic      dvd_release
##  Min.   : 1.100   Min.   :  1.00   Min.   :1990-12-19
##  1st Qu.: 5.600   1st Qu.: 43.00   1st Qu.:2009-08-11
##  Median : 6.400   Median : 56.00   Median :2012-10-09
##  Mean   : 6.311   Mean   : 55.75   Mean   :2011-08-07
##  3rd Qu.: 7.100   3rd Qu.: 69.00   3rd Qu.:2015-03-17
##  Max.   :10.000   Max.   :100.00   Max.   :2017-12-04
##  NA's   :733      NA's   :3389     NA's   :3133
##                  production
##  N/A                   :1707
##  Universal Pictures    : 317
##  IFC Films             : 180
##  Warner Bros. Pictures: 161
##  20th Century Fox      : 159
##  Magnolia Pictures     : 146
##  (Other)               :5798
##                                                              actors
##  N/A                                                            : 312
##                                                                 :   3
##  Bradley Cooper, Ed Helms, Zach Galifianakis, Justin Bartha:   3
##  Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott :   3
##  Adam Sandler, Andy Samberg, Selena Gomez, Kevin James     :   2
##  Adam Sandler, Kevin James, Chris Rock, David Spade        :   2
##  (Other)                                                        :8143
##    imdb_votes                     director     release_date
##  Min.   :      5.0   N/A             :  75   Min.   :1917-06-17
##  1st Qu.:    453.5   Steven Spielberg:  21   1st Qu.:2010-07-09
##  Median :   2920.0   Tyler Perry     :  16   Median :2013-12-13
##  Mean   :  34865.7   Ron Howard      :  14   Mean   :2012-01-19
##  3rd Qu.:  20943.5   Spike Lee       :  13   3rd Qu.:2016-02-04
##  Max.   :1827477.0   Ram Gopal Varma :  12   Max.   :2025-12-19
##  NA's   :733         (Other)         :8317   NA's   :185
##     runtime                         genre                        awards
##  N/A    : 619   Drama                 : 774   N/A                  :3223
##  90 min : 400   Documentary           : 583   1 nomination.        : 671
##  95 min : 228   Comedy                : 426   2 nominations.       : 292
##  100 min: 225   Comedy, Drama         : 389   1 win.               : 291
##  92 min : 213   Comedy, Drama, Romance: 289   1 win & 1 nomination.: 213
##  93 min : 208   Drama, Romance        : 249   3 nominations.       : 166
##  (Other):6575   (Other)               :5758   (Other)              :3612
##                     keywords        Budget           Box.Office.Gross
##                         :2087   Min.   :-1.997e+03   Min.   :        0
##  f-rated                : 108   1st Qu.: 0.000e+00   1st Qu.:        0
##  character-name-in-title:  11   Median : 1.374e+05   Median :        0
##  cgi-animation          :  10   Mean   : 3.458e+07   Mean   : 14907664
##  independent-film       :   9   3rd Qu.: 1.050e+07   3rd Qu.:   621738
##  \\N                    :   8   Max.   : 3.500e+10   Max.   :936662225
##  (Other)                :6235   NA's   :3            NA's   :259
```

Look at the time span of the movies here. Ranging from Choplin's film in 1917 till Avatar5 in 2025 that're still in plan.

```
# exclude keyword column here for tidiness
movies[,-15] %>%
```

```
  arrange(release_date) %>%
  top_n(10)
```

```
##       imdbid                                     title rating imdb_rating
## 1  tt0120338                                   Titanic  PG-13         7.7
## 2  tt0120915 Star Wars: Episode I - The Phantom Menace     PG         6.5
## 3  tt0468569                           The Dark Knight  PG-13         9.0
## 4  tt0499549                                    Avatar  PG-13         7.8
## 5  tt0848228                              The Avengers  PG-13         8.1
## 6  tt0369610                            Jurassic World  PG-13         7.0
## 7  tt2488496             Star Wars: The Force Awakens  PG-13         8.1
## 8  tt2277860                              Finding Dory     PG         7.4
## 9  tt3748528                                 Rogue One  PG-13         7.9
## 10 tt2771200                     Beauty and the Beast     PG         7.6
##    metacritic dvd_release                       production
## 1          74  2012-09-10                Paramount Pictures
## 2          51  2001-10-16                  20th Century Fox
## 3          82  2008-12-09 Warner Bros. Pictures/Legendary
## 4          83  2010-04-22                  20th Century Fox
## 5          69  2012-09-25              Walt Disney Pictures
## 6          59  2015-10-20                Universal Pictures
## 7          81  2016-04-05              Walt Disney Pictures
## 8          77  2016-11-15       Walt Disney Pictures/PIXAR
## 9          65  2017-04-04              Walt Disney Pictures
## 10         65  2017-06-06              Walt Disney Pictures
##                                                          actors
## 1          Leonardo DiCaprio, Kate Winslet, Billy Zane, Kathy Bates
## 2           Liam Neeson, Ewan McGregor, Natalie Portman, Jake Lloyd
## 3        Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine
## 4      Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang
## 5     Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth
## 6  Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio
## 7            Harrison Ford, Mark Hamill, Carrie Fisher, Adam Driver
## 8         Ellen DeGeneres, Albert Brooks, Ed O'Neill, Kaitlin Olson
## 9                Felicity Jones, Diego Luna, Alan Tudyk, Donnie Yen
## 10                   Emma Watson, Dan Stevens, Luke Evans, Josh Gad
##    imdb_votes                       director release_date runtime
## 1      847267                  James Cameron   1997-12-19 194 min
## 2      573249                   George Lucas   1999-05-19 136 min
## 3     1827477              Christopher Nolan   2008-07-18 152 min
## 4      949141                  James Cameron   2009-12-18 162 min
## 5     1046622                    Joss Whedon   2012-05-04 143 min
## 6      458076                Colin Trevorrow   2015-06-12 124 min
## 7      665521                    J.J. Abrams   2015-12-18 136 min
## 8      161168 Andrew Stanton, Angus MacLane   2016-06-17  97 min
## 9      329408                 Gareth Edwards   2016-12-16 133 min
## 10     115633                    Bill Condon   2017-03-17 129 min
##                           genre
## 1                 Drama, Romance
## 2     Action, Adventure, Fantasy
## 3           Action, Crime, Drama
## 4     Action, Adventure, Fantasy
## 5                 Action, Sci-Fi
## 6      Action, Adventure, Sci-Fi
```

```
## 7    Action, Adventure, Fantasy
## 8  Animation, Adventure, Comedy
## 9     Action, Adventure, Sci-Fi
## 10     Family, Fantasy, Musical
##                                                                  awards
## 1                  Won 11 Oscars. Another 110 wins & 74 nominations.
## 2           Nominated for 3 Oscars. Another 25 wins & 60 nominations.
## 3                  Won 2 Oscars. Another 151 wins & 154 nominations.
## 4                   Won 3 Oscars. Another 85 wins & 128 nominations.
## 5            Nominated for 1 Oscar. Another 37 wins & 78 nominations.
## 6                                          6 wins & 54 nominations.
## 7          Nominated for 5 Oscars. Another 51 wins & 115 nominations.
## 8  Nominated for 1 BAFTA Film Award. Another 9 wins & 40 nominations.
## 9           Nominated for 2 Oscars. Another 11 wins & 72 nominations.
## 10                                          2 wins & 3 nominations.
##      Budget Box.Office.Gross
## 1  2.00e+08        658672302
## 2  1.15e+08        474544677
## 3  1.85e+08        534858444
## 4  2.37e+08        760507625
## 5  2.20e+08        623357910
## 6  1.50e+08        652270625
## 7  2.45e+08        936662225
## 8  2.00e+08        486295561
## 9  2.00e+08        532177324
## 10 1.60e+08        504014165
```

```
movies[,-15] %>%
  arrange(desc(release_date)) %>%
  top_n(10)
```

```
##       imdbid                                    title rating imdb_rating
## 1  tt2771200                  Beauty and the Beast     PG         7.6
## 2  tt3748528                             Rogue One  PG-13         7.9
## 3  tt2277860                          Finding Dory     PG         7.4
## 4  tt2488496          Star Wars: The Force Awakens  PG-13         8.1
## 5  tt0369610                        Jurassic World  PG-13         7.0
## 6  tt0848228                          The Avengers  PG-13         8.1
## 7  tt0499549                                Avatar  PG-13         7.8
## 8  tt0468569                           The Dark Knight  PG-13         9.0
## 9  tt0120915 Star Wars: Episode I - The Phantom Menace     PG         6.5
## 10 tt0120338                                Titanic  PG-13         7.7
##    metacritic dvd_release                    production
## 1          65  2017-06-06            Walt Disney Pictures
## 2          65  2017-04-04            Walt Disney Pictures
## 3          77  2016-11-15      Walt Disney Pictures/PIXAR
## 4          81  2016-04-05            Walt Disney Pictures
## 5          59  2015-10-20              Universal Pictures
## 6          69  2012-09-25            Walt Disney Pictures
## 7          83  2010-04-22                20th Century Fox
## 8          82  2008-12-09 Warner Bros. Pictures/Legendary
## 9          51  2001-10-16                20th Century Fox
## 10         74  2012-09-10              Paramount Pictures
##                                                      actors
## 1                  Emma Watson, Dan Stevens, Luke Evans, Josh Gad
```

```
## 2                Felicity Jones, Diego Luna, Alan Tudyk, Donnie Yen
## 3         Ellen DeGeneres, Albert Brooks, Ed O'Neill, Kaitlin Olson
## 4            Harrison Ford, Mark Hamill, Carrie Fisher, Adam Driver
## 5  Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio
## 6     Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth
## 7      Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang
## 8        Christian Bale, Heath Ledger, Aaron Eckhart, Michael Caine
## 9           Liam Neeson, Ewan McGregor, Natalie Portman, Jake Lloyd
## 10         Leonardo DiCaprio, Kate Winslet, Billy Zane, Kathy Bates
##    imdb_votes                       director release_date runtime
## 1      115633                    Bill Condon   2017-03-17 129 min
## 2      329408                 Gareth Edwards   2016-12-16 133 min
## 3      161168 Andrew Stanton, Angus MacLane   2016-06-17  97 min
## 4      665521                    J.A. Abrams   2015-12-18 136 min
## 5      458076                Colin Trevorrow   2015-06-12 124 min
## 6     1046622                    Joss Whedon   2012-05-04 143 min
## 7      949141                  James Cameron   2009-12-18 162 min
## 8     1827477              Christopher Nolan   2008-07-18 152 min
## 9      573249                   George Lucas   1999-05-19 136 min
## 10     847267                  James Cameron   1997-12-19 194 min
##                           genre
## 1      Family, Fantasy, Musical
## 2     Action, Adventure, Sci-Fi
## 3  Animation, Adventure, Comedy
## 4    Action, Adventure, Fantasy
## 5     Action, Adventure, Sci-Fi
## 6                Action, Sci-Fi
## 7    Action, Adventure, Fantasy
## 8          Action, Crime, Drama
## 9    Action, Adventure, Fantasy
## 10               Drama, Romance
##                                                                  awards
## 1                                              2 wins & 3 nominations.
## 2           Nominated for 2 Oscars. Another 11 wins & 72 nominations.
## 3  Nominated for 1 BAFTA Film Award. Another 9 wins & 40 nominations.
## 4          Nominated for 5 Oscars. Another 51 wins & 115 nominations.
## 5                                            6 wins & 54 nominations.
## 6            Nominated for 1 Oscar. Another 37 wins & 78 nominations.
## 7                    Won 3 Oscars. Another 85 wins & 128 nominations.
## 8                   Won 2 Oscars. Another 151 wins & 154 nominations.
## 9           Nominated for 3 Oscars. Another 25 wins & 60 nominations.
## 10                  Won 11 Oscars. Another 110 wins & 74 nominations.
##      Budget Box.Office.Gross
## 1  1.60e+08        504014165
## 2  2.00e+08        532177324
## 3  2.00e+08        486295561
## 4  2.45e+08        936662225
## 5  1.50e+08        652270625
## 6  2.20e+08        623357910
## 7  2.37e+08        760507625
## 8  1.85e+08        534858444
## 9  1.15e+08        474544677
## 10 2.00e+08        658672302
```

Put the movies that're released after 2017 in a new set so we could predict the box office for them using historical box office value. (they have no budget and box office value in the dataframe now) To do that, first, we ned to seperate release date into new columns - year, month, and date.

```
movies$release.year <- as.numeric(format(movies$release_date, format = "%Y"))
movies$release.month <-as.numeric(format(movies$release_date, format = "%m"))
movies$release.day <-as.numeric(format(movies$release_date, format = "%d"))
```

```
movies.after.2017 <- movies %>%
  filter(release.year>2017)
```

```
movies.before <- movies %>%
  filter(release.year<=2017) %>%
  filter(Box.Office.Gross!=0) %>%
  filter(Budget!=0)
#head(movies.before)
```

Extract keyword column here for future use.

```
keyword <- as.character(movies.before$keywords)

movies.before <- subset(movies.before, select = -c(keywords))
```

Check and find that there're no more NAs and 0s in box office and budget. This is our main dataset for exploratory analysis and modeling building.

```
movies.before$roi <- (movies.before$Box.Office.Gross-movies.before$Budget)/movies.before$Budget
summary(movies.before)
```

```
##        imdbid                       title            rating
##  tt0042332:   1   Beauty and the Beast:   2   R        :1071
##  tt0046004:   1   Ben-Hur             :   2   PG-13    : 888
##  tt0047396:   1   Brothers            :   2   PG       : 345
##  tt0052618:   1   Cinderella          :   2   N/A      : 142
##  tt0054177:   1   Conan the Barbarian :   2   NOT RATED:  88
##  tt0058414:   1   Daddy's Home        :   2   G        :  47
##  (Other)  :2608   (Other)             :2602   (Other)  :  33
##   imdb_rating      metacritic     dvd_release
##  Min.   :1.600   Min.   :  1.0   Min.   :1991-12-19
##  1st Qu.:5.800   1st Qu.: 41.0   1st Qu.:2008-05-06
##  Median :6.500   Median : 53.0   Median :2011-04-19
##  Mean   :6.371   Mean   : 53.6   Mean   :2010-04-02
##  3rd Qu.:7.100   3rd Qu.: 66.0   3rd Qu.:2014-04-08
##  Max.   :9.900   Max.   :100.0   Max.   :2017-08-15
##  NA's   :17      NA's   :333     NA's   :201
##                  production
##  Universal Pictures   : 263
##  Warner Bros. Pictures: 140
##  20th Century Fox     : 134
##  Sony Pictures        : 109
##  Paramount Pictures   :  96
##  N/A                  :  69
##  (Other)              :1803
##                                                              actors
##  N/A                                                            :   7
##  Bradley Cooper, Ed Helms, Zach Galifianakis, Justin Bartha     :   3
##  Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott      :   3
```

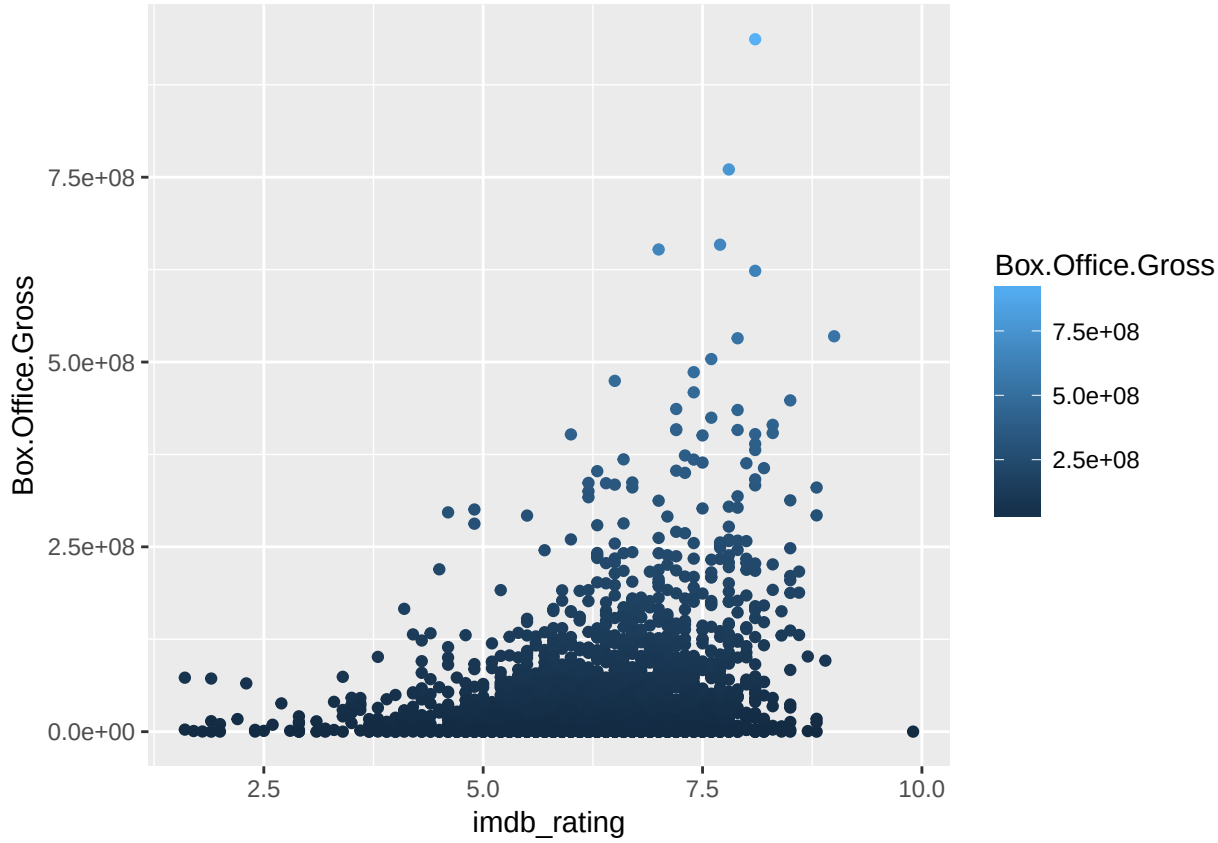
```
##  Adam Sandler, Andy Samberg, Selena Gomez, Kevin James       :   2
##  Adam Sandler, Kevin James, Chris Rock, David Spade          :   2
##  Ben Stiller, Chris Rock, David Schwimmer, Jada Pinkett Smith:   2
##  (Other)                                                     :2595
##    imdb_votes                   director     release_date
##  Min.   :     17   Steven Spielberg :  18   Min.   :1950-03-04
##  1st Qu.:   9142   Ron Howard       :  13   1st Qu.:2008-02-02
##  Median :  38340   Steven Soderbergh:  10   Median :2011-05-02
##  Mean   :  92852   Dennis Dugan     :   9   Mean   :2009-03-05
##  3rd Qu.: 108550   Robert Zemeckis  :   9   3rd Qu.:2014-05-30
##  Max.   :1827477   Spike Lee        :   9   Max.   :2017-08-18
##  NA's   :17        (Other)          :2546
##     runtime                        genre                         awards
##  90 min :  86   Drama                 : 145   N/A                   : 429
##  100 min:  75   Comedy, Drama         : 110   1 nomination.         : 198
##  98 min :  73   Comedy, Drama, Romance: 109   2 nominations.        : 104
##  97 min :  72   Comedy                : 107   1 win.                :  78
##  104 min:  68   Comedy, Romance       :  91   3 nominations.        :  72
##  92 min :  68   Drama, Romance        :  88   1 win & 1 nomination.:  57
##  (Other):2172   (Other)               :1964   (Other)               :1676
##      Budget          Box.Office.Gross     release.year  release.month
##  Min.   :4.400e+01   Min.   :      335   Min.   :1950   Min.   : 1.000
##  1st Qu.:7.000e+06   1st Qu.:  1243848   1st Qu.:2008   1st Qu.: 4.000
##  Median :2.000e+07   Median : 18912638   Median :2011   Median : 7.000
##  Mean   :6.418e+07   Mean   : 46815764   Mean   :2009   Mean   : 6.567
##  3rd Qu.:5.000e+07   3rd Qu.: 58382184   3rd Qu.:2014   3rd Qu.:10.000
##  Max.   :3.000e+10   Max.   :936662225   Max.   :2017   Max.   :12.000
##
##   release.day         roi
##  Min.   : 1.00   Min.   :  -1.000
##  1st Qu.: 9.00   1st Qu.:  -0.804
##  Median :17.00   Median :  -0.183
##  Mean   :16.21   Mean   :   5.976
##  3rd Qu.:23.00   3rd Qu.:   0.835
##  Max.   :31.00   Max.   :7193.587
##
```

EXPLORATORY ANALYSIS

1. Does good imdb rating positively correlated with good box office?

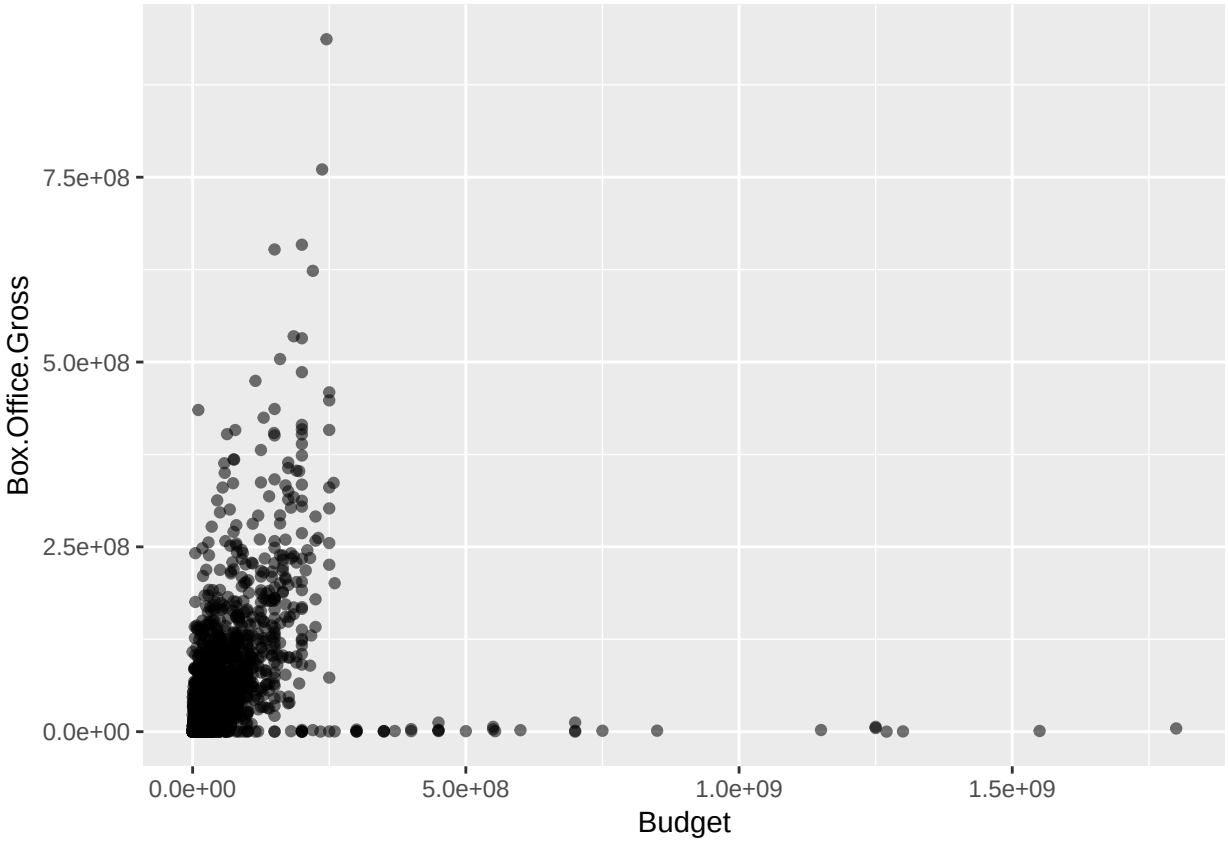
Yes. Movies that have achieved high box office gross tend to have higher ratings. But movies that have higher ratings don't necessarily have high box office gross.

```
movies.before %>%
  ggplot(aes(x=imdb_rating,y=Box.Office.Gross,col=Box.Office.Gross)) +
  geom_point()
```



2. Relationship of budget and box office gross:

```
# 2 outliers with extremly high busget
movies.before %>%
  filter(Budget<5000000000) %>%
  ggplot(aes(x=Budget,y=Box.Office.Gross,alpha=0.5))+
  geom_point(show.legend = FALSE)
```



Observe which movies have high budgets and low box office. Most of them win lots of awards and nominations. Probably more artistic and only appeal to a niche market than a blockbuster.

```
movies.before %>%
  filter(Budget>250000000) %>%
  select(title,imdb_rating,release_date,production,actors,director,genre,awards,Budget,Box.Office.Gross
  top_n(10)
```

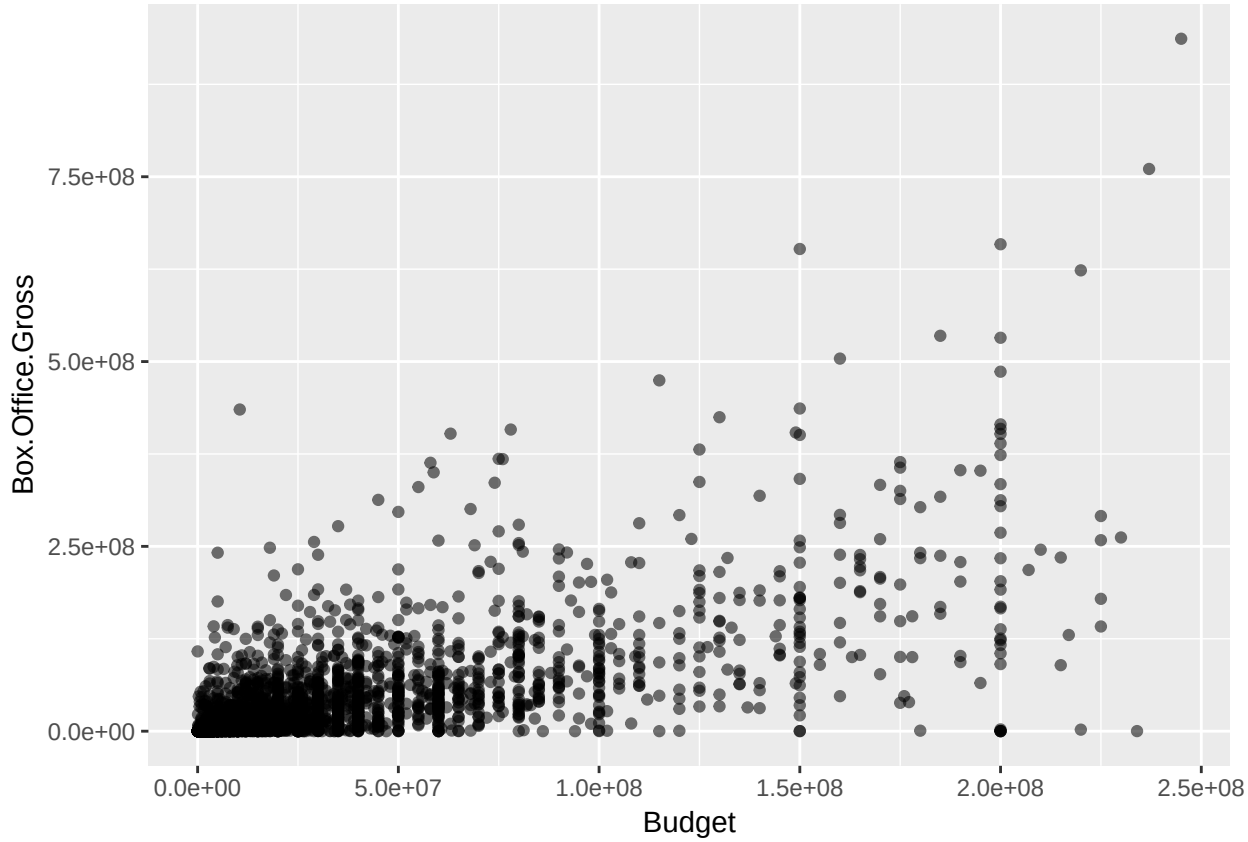
```
##                       title imdb_rating release_date
## 1                   Tangled         7.8   2010-11-24
## 2            Bajirao Mastani         7.2   2015-12-18
## 3                     Dangal         8.8   2016-12-21
## 4               Spider-Man 3         6.2   2007-05-04
## 5               Jodhaa Akbar         7.6   2008-02-15
## 6                   3 Idiots         8.4   2009-12-25
## 7  Zindagi Na Milegi Dobara         8.1   2011-07-15
## 8                     Barfi!         8.1   2012-09-14
## 9                  Dabangg 2         4.9   2012-12-21
## 10                 Singam 2         6.3   2013-07-04
##                 production
## 1    Walt Disney Pictures
## 2                SLB Films
## 3  Aamir Khan Productions
## 4           Sony Pictures
## 5      UTV Communications
## 6             Big Pictures
## 7      Eros International
```

```
## 8      UTV Communications
## 9      Eros Entertainment
## 10                 ATMUS
##                                                              actors
## 1             Mandy Moore, Zachary Levi, Donna Murphy, Ron Perlman
## 2     Ranveer Singh, Priyanka Chopra, Deepika Padukone, Tanvi Azmi
## 3    Aamir Khan, Sakshi Tanwar, Fatima Sana Shaikh, Sanya Malhotra
## 4  Tobey Maguire, Kirsten Dunst, James Franco, Thomas Haden Church
## 5  Hrithik Roshan, Aishwarya Rai Bachchan, Sonu Sood, Poonam Sinha
## 6         Aamir Khan, Madhavan, Sharman Joshi, Kareena Kapoor Khan
## 7   Hrithik Roshan, Katrina Kaif, Naseeruddin Shah, Kalki Koechlin
## 8    Ranbir Kapoor, Priyanka Chopra, Ileana D'Cruz, Saurabh Shukla
## 9           Salman Khan, Sonakshi Sinha, Prakash Raj, Vinod Khanna
## 10                  Suriya, Anushka Shetty, Hansika Motwani, Vivek
##                       director                        genre
## 1  Nathan Greno, Byron Howard Animation, Adventure, Comedy
## 2       Sanjay Leela Bhansali       Action, Drama, History
## 3                Nitesh Tiwari     Action, Biography, Drama
## 4                    Sam Raimi            Action, Adventure
## 5           Ashutosh Gowariker Action, Adventure, Biography
## 6              Rajkumar Hirani     Adventure, Comedy, Drama
## 7                  Zoya Akhtar     Adventure, Comedy, Drama
## 8                  Anurag Basu     Adventure, Comedy, Drama
## 9                  Arbaaz Khan        Action, Comedy, Drama
## 10                       Hari             Action, Thriller
##                                                                awards
## 1             Nominated for 1 Oscar. Another 9 wins & 40 nominations.
## 2                                          29 wins & 11 nominations.
## 3                                                          3 wins.
## 4  Nominated for 1 BAFTA Film Award. Another 3 wins & 32 nominations.
## 5                                          22 wins & 25 nominations.
## 6                                          25 wins & 13 nominations.
## 7                                          19 wins & 16 nominations.
## 8                                          35 wins & 21 nominations.
## 9                                            6 wins & 6 nominations.
## 10                                                   8 nominations.
##      Budget Box.Office.Gross        roi
## 1  2.60e+08        200821936 -0.2276079
## 2  1.25e+09          6557047 -0.9947544
## 3  7.00e+08         12391761 -0.9822975
## 4  2.58e+08        336530303  0.3043810
## 5  4.00e+08          3440718 -0.9913982
## 6  5.50e+08          6532908 -0.9881220
## 7  5.50e+08          3076226 -0.9944069
## 8  3.00e+08          2804874 -0.9906504
## 9  4.50e+08          2519190 -0.9944018
## 10 4.50e+08         12331200 -0.9725973
```

When the budget is within reasonable amount, higher budget may bring more box office though not always the case.

```
#remove outliers to make the plot pattern clearer
movies.before %>%
  filter(Budget<250000000) %>%
  ggplot(aes(x=Budget,y=Box.Office.Gross,alpha=0.5))+
```

```
  geom_point(show.legend = FALSE)
```

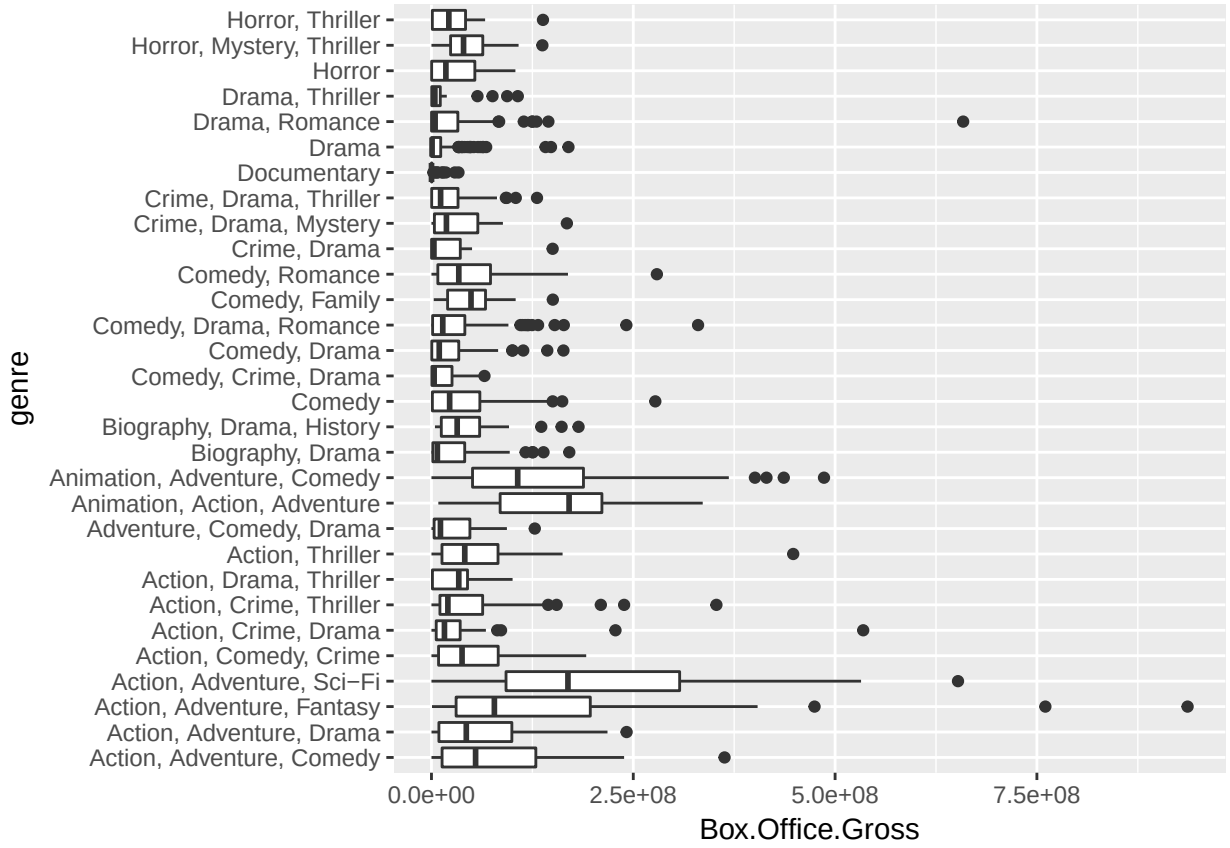


3. Do certain genre lend themselves to higher return?

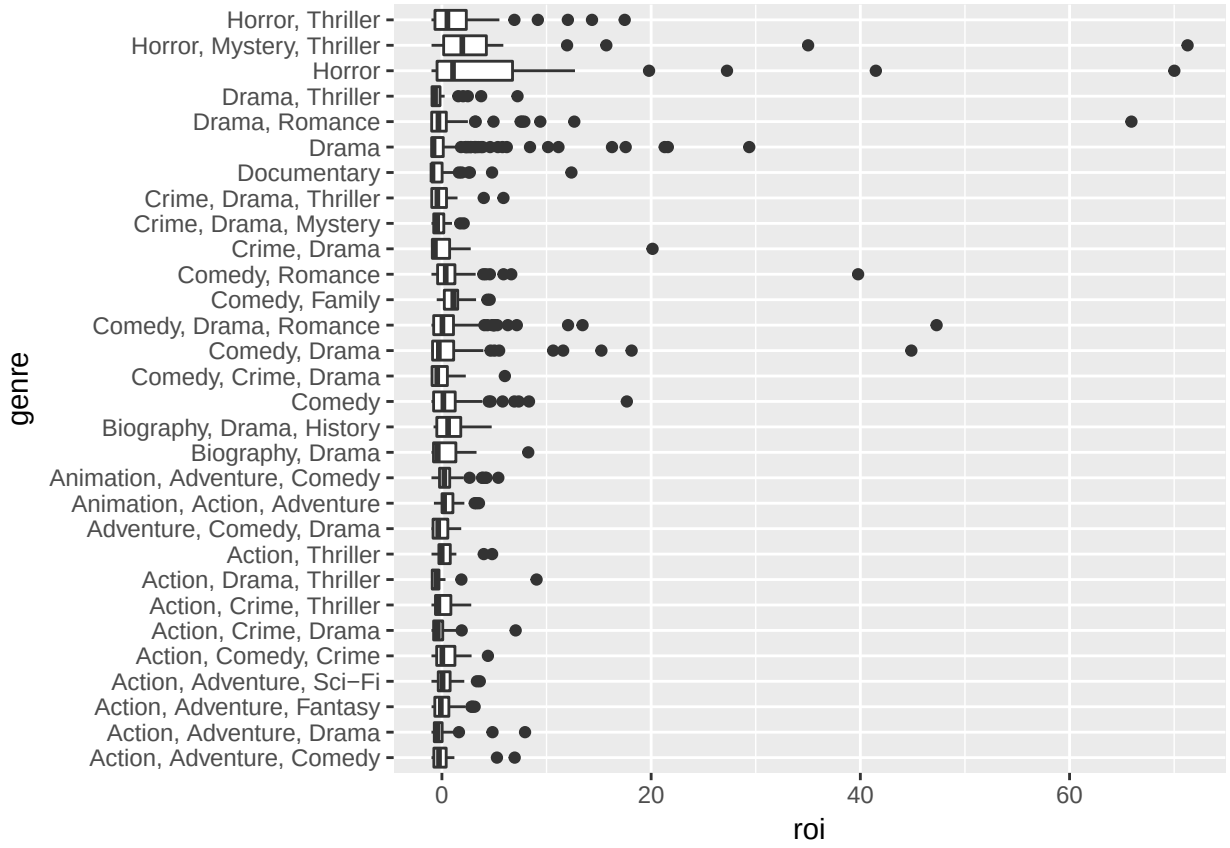
```
genre <- movies.before %>%
  count(genre) %>%
  arrange(desc(n))
head(genre)
```

```
## # A tibble: 6 x 2
##   genre                      n
##   <fct>                  <int>
## 1 Drama                    145
## 2 Comedy, Drama            110
## 3 Comedy, Drama, Romance   109
## 4 Comedy                   107
## 5 Comedy, Romance           91
## 6 Drama, Romance            88
```

```
movies.before %>%
  inner_join(genre[1:30,],by="genre") %>%
  ggplot(aes(x=genre,y=Box.Office.Gross))+
  geom_boxplot()+
  coord_flip()
```




```
#remove an outlier that has over 6000 roi in horror,mystery,thriller genre
movies.before %>%
  inner_join(genre[1:30,],by="genre") %>%
  filter(roi<100) %>%
  ggplot(aes(x=genre,y=roi))+
  geom_boxplot()+
  coord_flip()
```



Intersting and important Findings:

- Although there're more than 300 segmented genres, it's clear in the plot that they could be integrated into three main categories: comedy & romance & drama, horror & thriller & mystery, and action & adverture & crime. Sci-Fi and Documentary are in smaller amount so these two genres didn't really stand out.
- In terms of box office gross, action > comedy > thriller. It makes sense that people like watching action movies because it's exciting, and thriller is less relatable to the general public.
- In terms of roi, thriller > comedy > action. Action movie needs more resources to shoot and make the setting, while thriller is generally less cost-consuming.

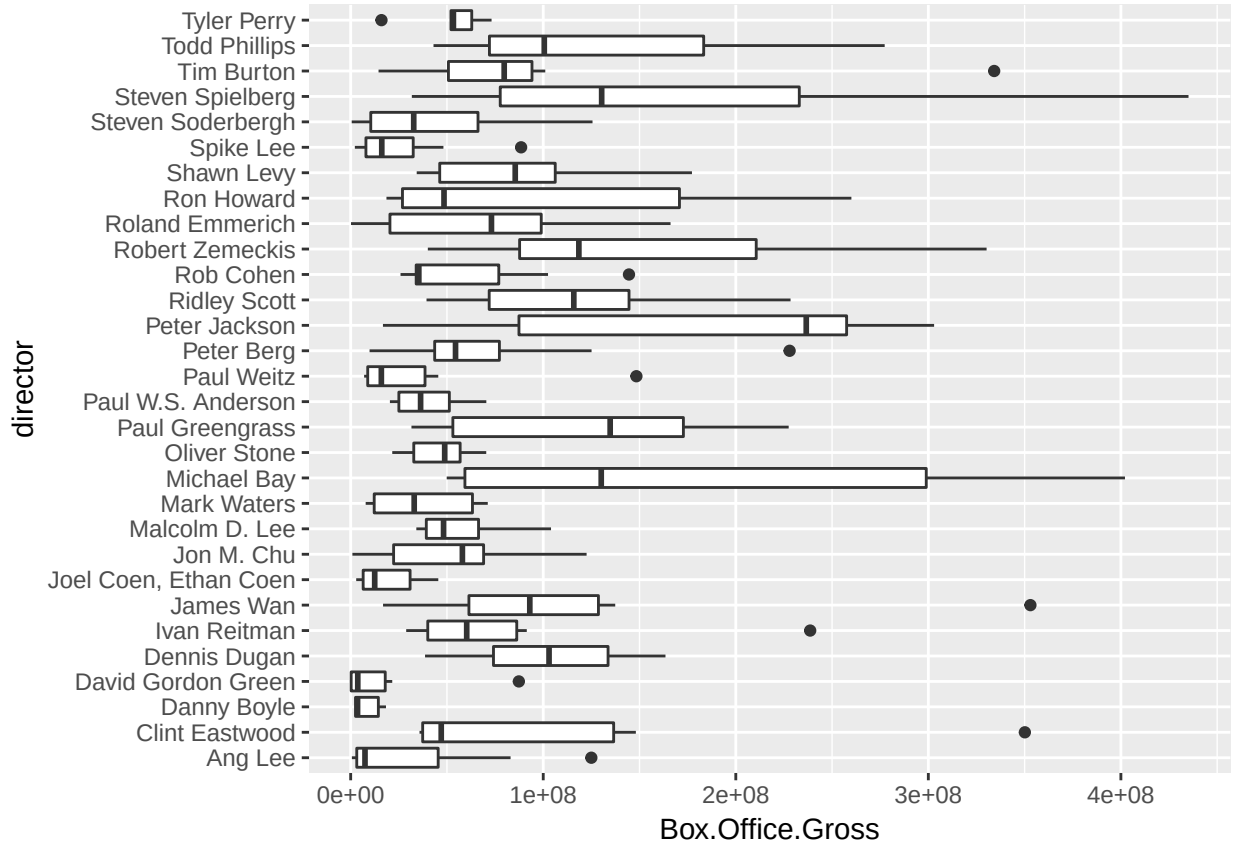
4. Relationship of director and box office

```
director <- movies.before %>%
  count(director,sort=TRUE)
head(director)
```

```
## # A tibble: 6 x 2
##   director              n
##   <fct>             <int>
## 1 Steven Spielberg     18
## 2 Ron Howard           13
## 3 Steven Soderbergh    10
## 4 Dennis Dugan          9
## 5 Robert Zemeckis       9
## 6 Spike Lee             9
```

This plot is sort by number of movies each director has in the dataset, because we're more interested in well-known directors who have produced many award-winning movies. But it also means directors who has less movies might be left out in this plot.

```
movies.before %>%
  inner_join(director[1:30,],by="director") %>%
  ggplot(aes(x=director,y=Box.Office.Gross))+
  geom_boxplot()+
  coord_flip()
```



Observe what type of movies those directors with higher box office shot. (more of a genre issue or director issue.)

Jurassic Park series, ET, Schindler's list, AI, Indiana Jones. Adventure, sci-fi, drama, war, history, action, biography

```
movies.before %>%
  filter(director=="Steven Spielberg") %>%
  select(title,genre,Box.Office.Gross,roi)
```

```
##                                    title
## 1                           Jurassic Park
## 2                                 Lincoln
## 3                 The Adventures of Tintin
## 4                               War Horse
## 5                          Bridge of Spies
## 6                                 The BFG
## 7                                    1941
```

```
## 8                              Raiders of the Lost Ark
## 9                            E.T. the Extra-Terrestrial
## 10                                             Always
## 11                                   Schindler's List
## 12                         The Lost World: Jurassic Park
## 13                                   Saving Private Ryan
## 14                          A.I. Artificial Intelligence
## 15                                   Catch Me If You Can
## 16                                         The Terminal
## 17 Indiana Jones and the Kingdom of the Crystal Skull
## 18                                     War of the Worlds
##                             genre Box.Office.Gross         roi
## 1   Adventure, Sci-Fi, Thriller        402453882  5.38815686
## 2     Biography, Drama, History        182207973  1.80319958
## 3  Animation, Action, Adventure         77591831 -0.42524570
## 4                    Drama, War         79884879  0.21037695
## 5      Drama, History, Thriller         72313754  0.80784385
## 6    Adventure, Family, Fantasy         55483770 -0.60368736
## 7           Action, Comedy, War         31755742 -0.09269309
## 8             Action, Adventure        248159971 12.78666506
## 9                Family, Sci-Fi        435110554 40.43910038
## 10             Fantasy, Romance         43858790  0.41479968
## 11    Biography, Drama, History         96067179  3.36668995
## 12    Action, Adventure, Sci-Fi        229086679  2.13817368
## 13                   Drama, War        216540909  2.09344156
## 14     Adventure, Drama, Sci-Fi         78616689 -0.21383311
## 15      Biography, Crime, Drama        164615351  2.16567983
## 16       Comedy, Drama, Romance         77872883  0.29788138
## 17   Action, Adventure, Fantasy        317101119  0.71406010
## 18  Adventure, Sci-Fi, Thriller        234280354  0.77485117
```

Hobbit series and King Kong. Fantasy and adventure.

```
movies.before %>%
  filter(director=="Peter Jackson") %>%
  select(title,genre,Box.Office.Gross,roi)
```

```
##                                     title                    genre
## 1                         The Lovely Bones Drama, Fantasy, Thriller
## 2         The Hobbit: An Unexpected Journey       Adventure, Fantasy
## 3       The Hobbit: The Desolation of Smaug       Adventure, Fantasy
## 4 The Hobbit: The Battle of the Five Armies       Adventure, Fantasy
## 5                           The Frighteners  Comedy, Fantasy, Horror
## 6                                 King Kong Action, Adventure, Drama
##   Box.Office.Gross         roi
## 1         43818839 -0.32586402
## 2        303003568  0.68335316
## 3        258366855  0.14829713
## 4        255119788  0.02047915
## 5         16759216 -0.44135947
## 6        218080025  0.05352669
```

Transformers series, action.

```
movies.before %>%
  filter(director=="Michael Bay") %>%
```

```
  select(title,genre,Box.Office.Gross,roi)
```

```
##                                 title                     genre
## 1      Transformers: Dark of the Moon Action, Adventure, Sci-Fi
## 2                         Pain & Gain      Comedy, Crime, Drama
## 3     Transformers: Age of Extinction Action, Adventure, Sci-Fi
## 4       Transformers: The Last Knight Action, Adventure, Sci-Fi
## 5                           13 Hours    Action, Drama, History
## 6                           Bad Boys     Action, Comedy, Crime
## 7 Transformers: Revenge of the Fallen Action, Adventure, Sci-Fi
##   Box.Office.Gross         roi
## 1        352390543  0.80713099
## 2         49875291  0.91828042
## 3        245439076  0.16875750
## 4        130120862 -0.40036469
## 5         52853219  0.05706438
## 6         65807024  2.46352758
## 7        402111870  1.01055935
```

Titanic, Avatar, terminator. Action, sci-fi, adventure.

```
movies.before %>%
  filter(director=="James Cameron") %>%
  select(title,genre,Box.Office.Gross,roi)
```

```
##                        title                     genre Box.Office.Gross
## 1                    Titanic             Drama, Romance        658672302
## 2 Terminator 2: Judgment Day   Action, Sci-Fi, Thriller        204843350
## 3                  True Lies   Action, Comedy, Thriller        146282411
## 4                     Avatar Action, Adventure, Fantasy        760507625
##        roi
## 1 2.293362
## 2 1.008268
## 3 0.272021
## 4 2.208893
```

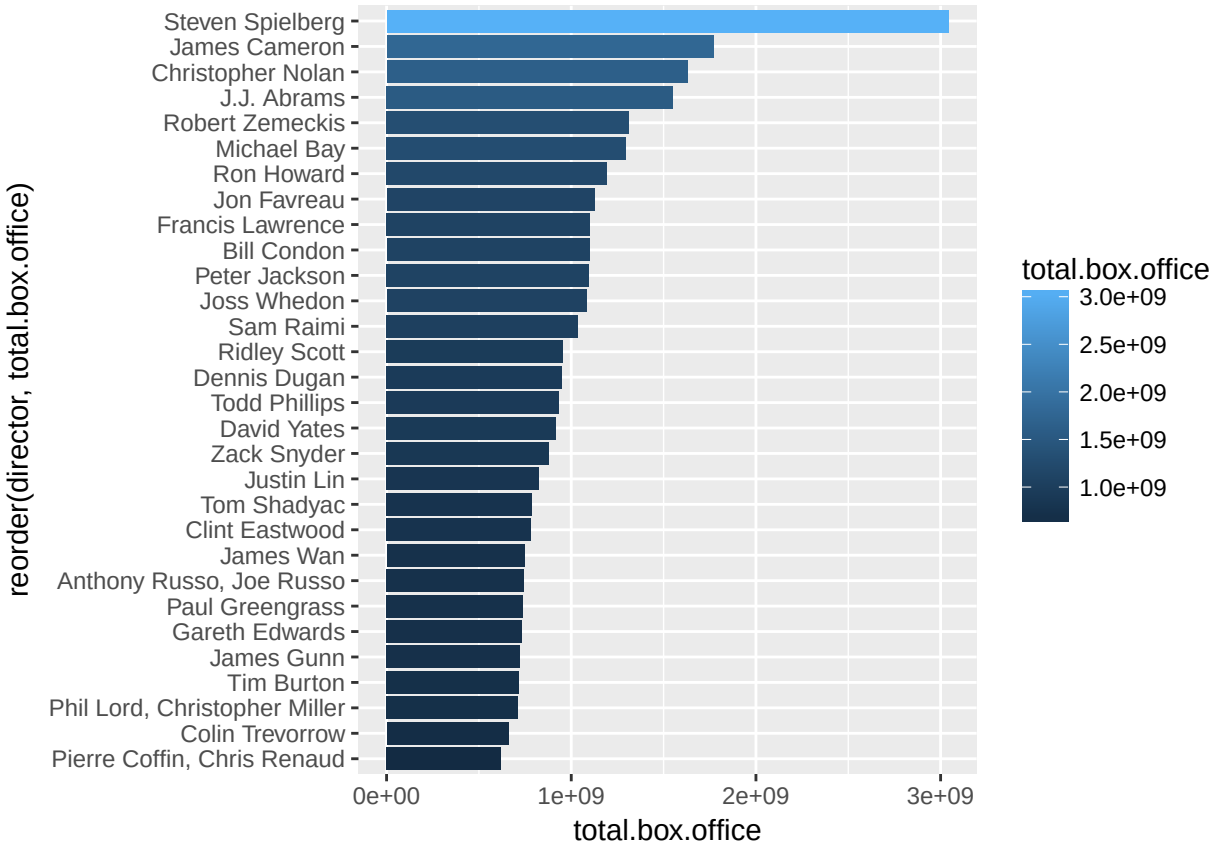
Sum of box office of every director:

```
movies.before %>%
  group_by(director) %>%
  summarize(total.box.office=sum(Box.Office.Gross)) %>%
  arrange(desc(total.box.office)) %>%
  top_n(10)
```

```
## # A tibble: 10 x 2
##    director          total.box.office
##    <fct>                        <dbl>
##  1 Steven Spielberg       3043002309.
##  2 James Cameron          1770305688.
##  3 Christopher Nolan      1629016219.
##  4 J.J. Abrams            1550175084.
##  5 Robert Zemeckis        1310841644.
##  6 Michael Bay            1298597885.
##  7 Ron Howard             1195612862.
##  8 Jon Favreau            1125727442.
##  9 Francis Lawrence       1102237551.
```

```
## 10 Bill Condon             1098592039.
```

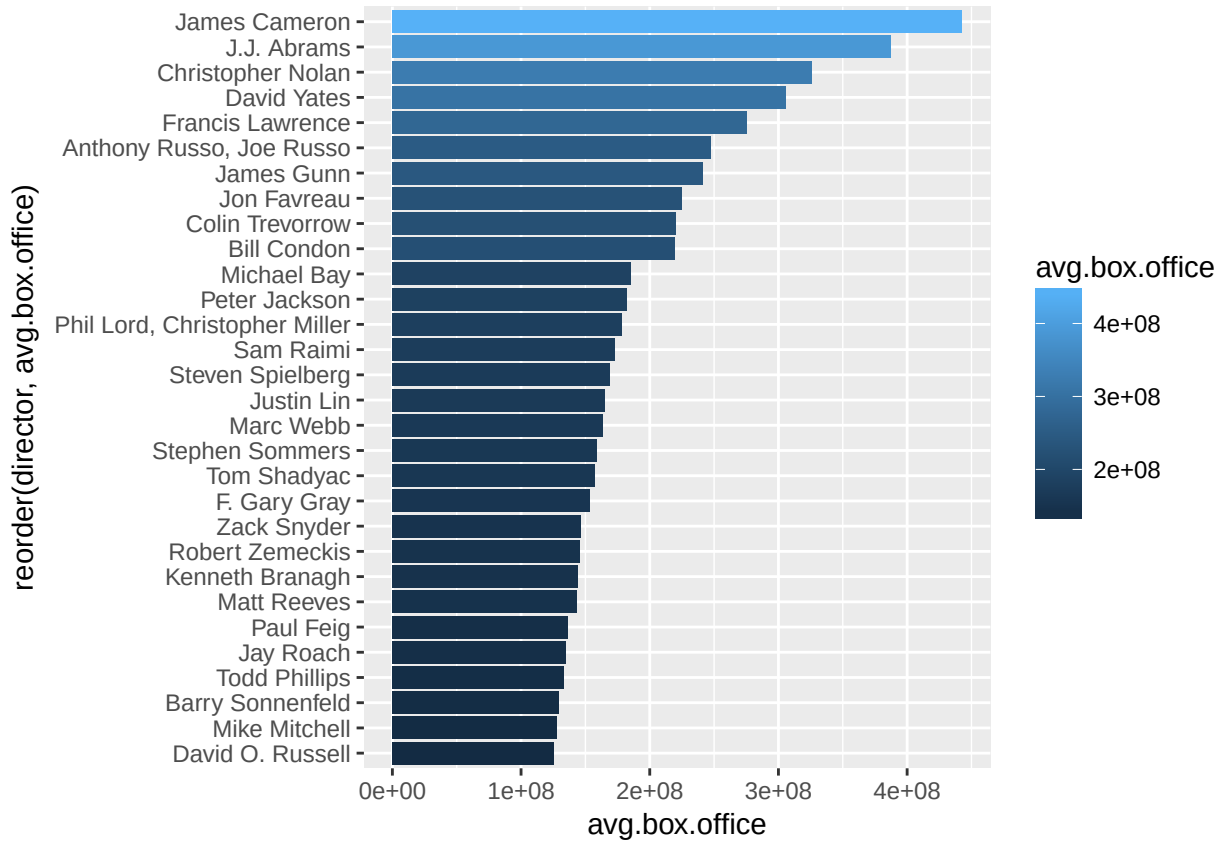
```
movies.before %>%
  group_by(director) %>%
  summarize(total.box.office=sum(Box.Office.Gross)) %>%
  arrange(desc(total.box.office)) %>%
  top_n(30) %>%
  ggplot(aes(x=reorder(director,total.box.office),y=total.box.office,fill=total.box.office))+
  geom_col()+
  coord_flip()
```



Average box office of for directors who have more than 3 movies in this dataset:

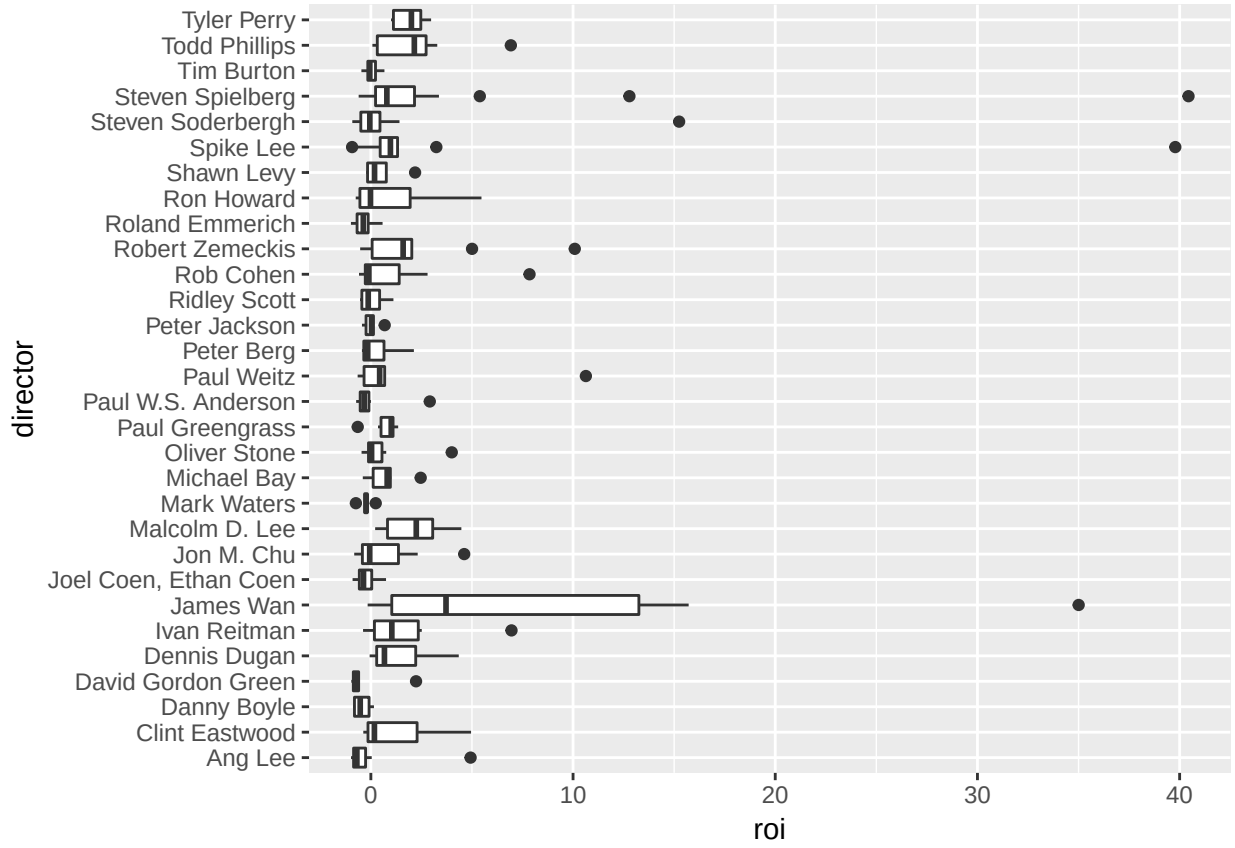
James Cameron comes on top this time. He's not as productive as spielberg, but he's definitely the lucrative.

```
#set filter because I want to eliminate directors who only have limited movie so his box office perform
movies.before %>%
  group_by(director) %>%
  filter(n()>=3) %>%
  summarize(avg.box.office=sum(Box.Office.Gross)/n()) %>%
  arrange(desc(avg.box.office)) %>%
  top_n(30) %>%
  ggplot(aes(x=reorder(director,avg.box.office),y=avg.box.office,fill=avg.box.office))+
  geom_col()+
  coord_flip()
```

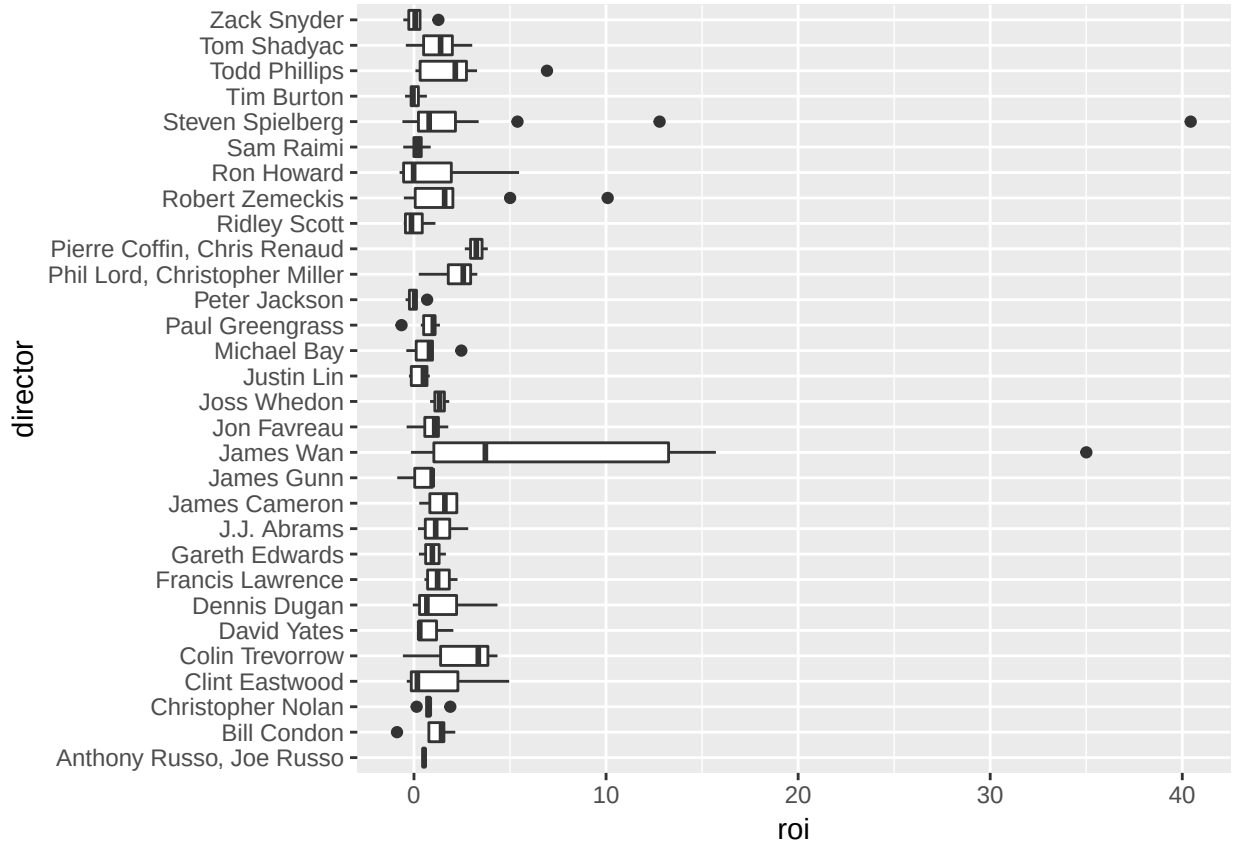


Compare roi:

```
# this is based on the number plot
movies.before %>%
  inner_join(director[1:30,],by="director") %>%
  ggplot(aes(x=director,y=roi))+
  geom_boxplot()+
  coord_flip()
```

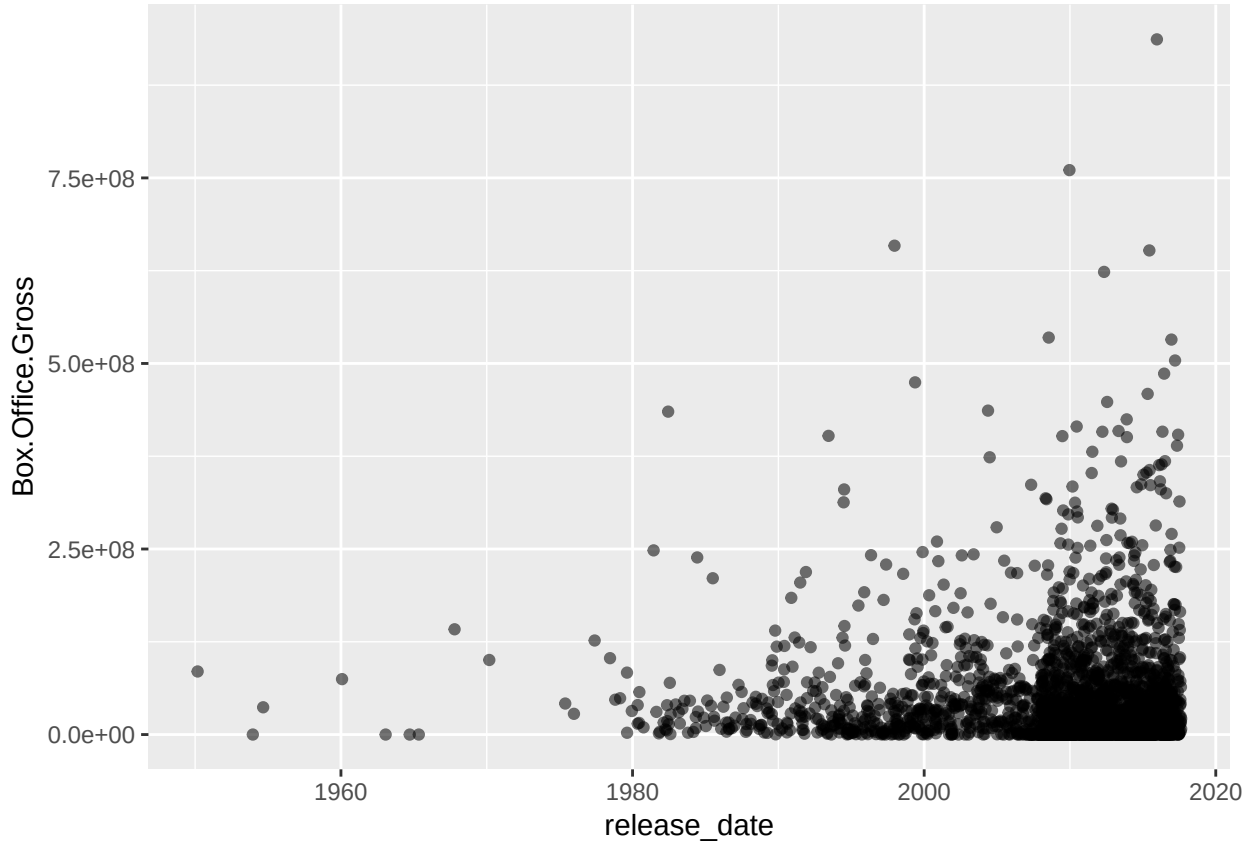



```
# this is based on the total.box.office plot
total.box.office <- movies.before %>%
  group_by(director) %>%
  summarize(total.box.office=sum(Box.Office.Gross)) %>%
  arrange(desc(total.box.office))

movies.before %>%
  inner_join(total.box.office[1:30,],by="director") %>%
  ggplot(aes(x=director,y=roi))+
  geom_boxplot()+
  coord_flip()
```



5. timeline

```
movies.before %>%
  ggplot(aes(x=release_date,y=Box.Office.Gross,alpha=0.3))+
  geom_point(show.legend = FALSE)
```



see how many movies are released every year.

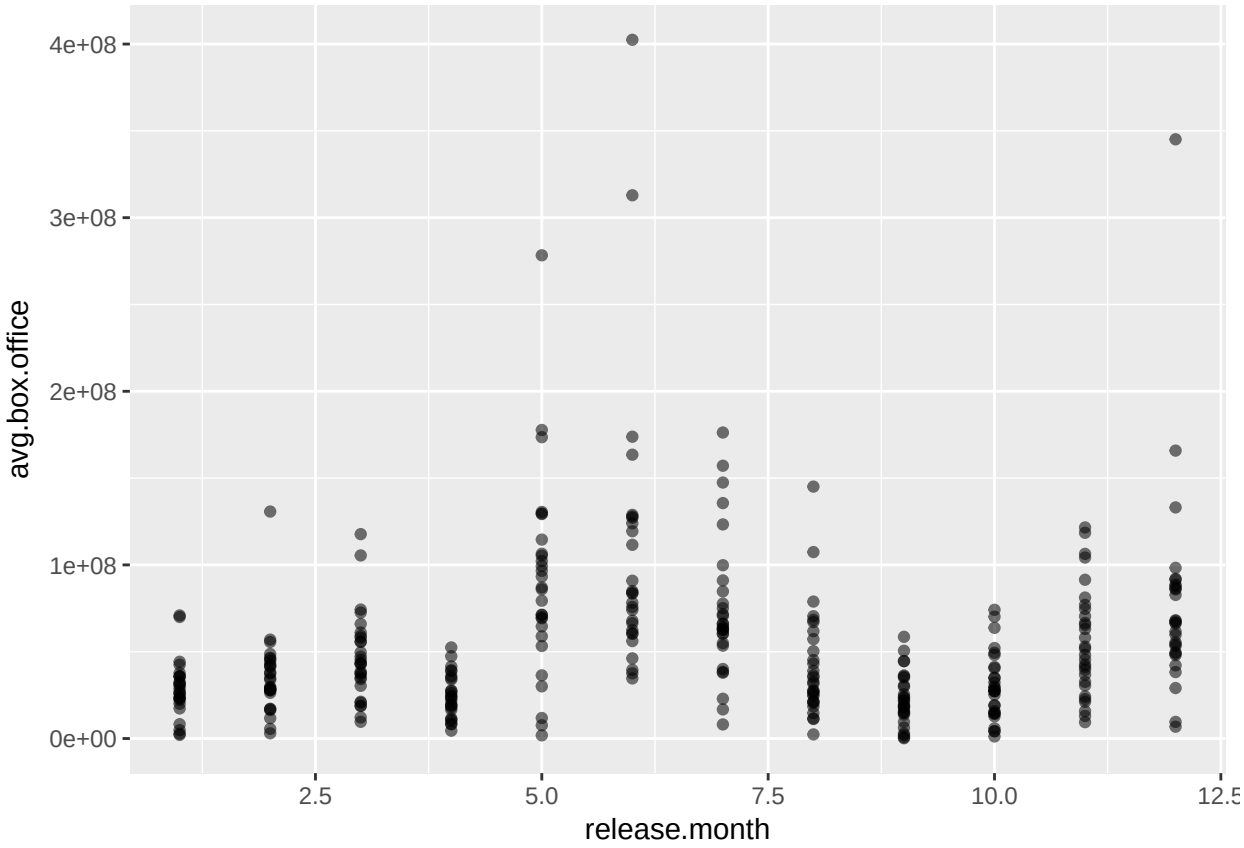
```
movies.before %>%
  group_by(release.year) %>%
  summarize(count=n()) %>%
  top_n(10)
```

```
## # A tibble: 10 x 2
##    release.year count
##           <dbl> <int>
##  1        2008.   213
##  2        2009.   203
##  3        2010.   188
##  4        2011.   211
##  5        2012.   216
##  6        2013.   212
##  7        2014.   217
##  8        2015.   204
##  9        2016.   218
## 10        2017.    99
```

To observe how time of the year influence box office, we choose year between 1988-2017 since it has more data. The original idea is to plot every year as a line with points in 12 months (avg.box.office). But I ran into difficulties here. I'll get back to it later.

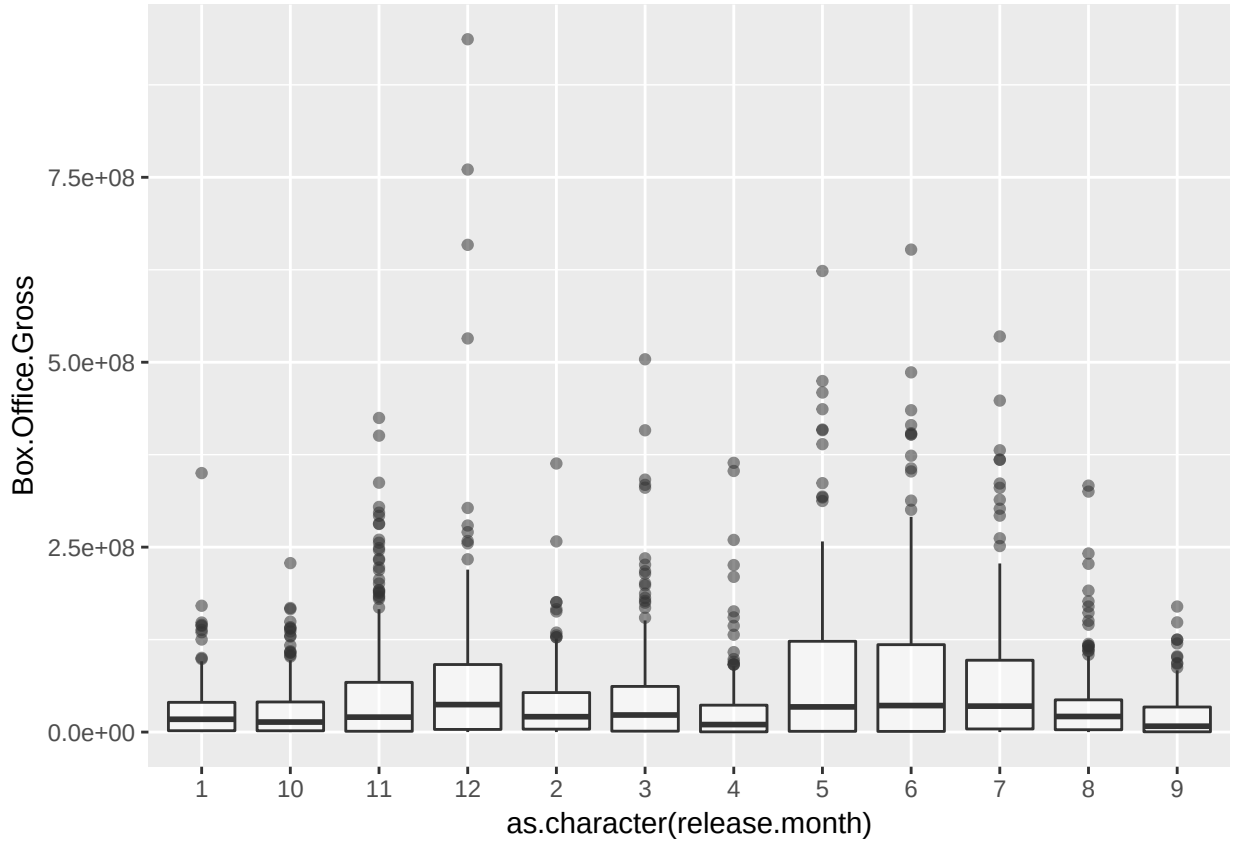
```
movies.before %>%
  filter(release.year>=1988) %>%
  group_by(release.year,release.month) %>%
```

```
  summarize(count=n(),avg.box.office=sum(Box.Office.Gross)/count) %>%
  ggplot(aes(x=release.month,y=avg.box.office,alpha=0.3))+
  geom_point(show.legend = FALSE)
```



This time I don't use average monthly box office, I put the box office of every movie here. The results are similar: movies released in May, June, July (summer time) and December tends to have more box office. It makes sense.

```
movies.before %>%
  ggplot(aes(x=as.character(release.month),y=Box.Office.Gross,alpha=0.1))+
  geom_boxplot(show.legend = FALSE)
```



6. Do remakes (adaptions), tent-poles, and sequels perform differently?

Fitst step, find sequels. Sort the titles by alphabetical orders and you'll find the sequels: 21/22 Jump Street, A Huanted House, Alvin and the Chipmunks, American Pie... But it's time-consuming. A clever way would be to group by director, since a series of movies are always filmed by the same director.

```
movies.before %>%
  group_by(title) %>%
  arrange(title) %>%
  top_n(10)
```

```
## # A tibble: 2,614 x 20
## # Groups:   title [2,595]
##    imdbid  title    rating imdb_rating metacritic dvd_release production
##    <fct>   <fct>    <fct>        <dbl>      <dbl> <date>      <fct>
##  1 tt1179~ 10 Clov~ PG-13         7.20        76. 2016-06-14  Bad Robot P~
##  2 tt3453~ 10 Days~ R             6.60       NaN  NA          N/A
##  3 tt0443~ 10,000 ~ PG-13         5.10        34. 2008-06-24  Warner Bros~
##  4 tt0488~ 12       PG-13         7.80        72. 2009-07-14  Sony Pictur~
##  5 tt1160~ 12 Roun~ PG-13         5.60        38. 2009-06-30  20th Centur~
##  6 tt2024~ 12 Year~ R             8.10        96. 2014-03-04  Fox Searchl~
##  7 tt1542~ 127 Hou~ R             7.60        82. 2011-03-01  Fox Searchl~
##  8 tt0337~ 13 Goin~ PG-13         6.10        57. 2006-02-07  Sony Pictur~
##  9 tt4172~ 13 Hours R             7.30        48. 2016-06-07  Paramount P~
## 10 tt2059~ 13 Sins  R             6.30        44. 2014-06-16  Radius-TWC
## # ... with 2,604 more rows, and 13 more variables: actors <fct>,
## #   imdb_votes <dbl>, director <fct>, release_date <date>, runtime <fct>,
```

```
## #   genre <fct>, awards <fct>, Budget <dbl>, Box.Office.Gross <dbl>,
## #   release.year <dbl>, release.month <dbl>, release.day <dbl>, roi <dbl>
```

I chose the top 50 director on the director dataframe sorted by the number of their films on the dataset, and sort it so it's easier to spot the sequels, especailly the well-known ones.

I saw Twilight Saga, X-Men, Fast and Furious, Transformers, the Hobbit, Back to the future, Spider Man, Madea series, the hangover, Star war series, Jurassic Park series here.

```
movies.before %>%
  inner_join(director[1:50,],by="director") %>%
  arrange(director) %>%
  top_n(10)
```

```
##       imdbid                                                title rating
## 1  tt0107290                                       Jurassic Park  PG-13
## 2  tt0443272                                             Lincoln  PG-13
## 3  tt0983193                          The Adventures of Tintin     PG
## 4  tt1568911                                           War Horse  PG-13
## 5  tt3682448                                    Bridge of Spies  PG-13
## 6  tt3691740                                            The BFG     PG
## 7  tt0078723                                               1941     PG
## 8  tt0082971                             Raiders of the Lost Ark     PG
## 9  tt0083866                        E.T. the Extra-Terrestrial     PG
## 10 tt0096794                                             Always     PG
## 11 tt0108052                                   Schindler's List      R
## 12 tt0119567                      The Lost World: Jurassic Park  PG-13
## 13 tt0120815                                Saving Private Ryan      R
## 14 tt0212720                       A.I. Artificial Intelligence  PG-13
## 15 tt0264464                                 Catch Me If You Can  PG-13
## 16 tt0362227                                       The Terminal  PG-13
## 17 tt0367882 Indiana Jones and the Kingdom of the Crystal Skull  PG-13
## 18 tt0407304                                 War of the Worlds  PG-13
##    imdb_rating metacritic dvd_release              production
## 1          8.1         68  2000-10-10   Universal City Studios
## 2          7.4         86  2013-03-26      Dreamworks Pictures
## 3          7.4         68  2012-03-13               Paramount
## 4          7.2         72  2012-04-03     Walt Disney Pictures
## 5          7.6         81  2016-02-02      Dreamworks Pictures
## 6          6.4         66  2016-11-29     Walt Disney Pictures
## 7          5.9         34  1999-03-23 MCA Universal Home Video
## 8          8.5         85  2003-10-21       Paramount Pictures
## 9          7.9         91  2002-10-22       Universal Pictures
## 10         6.4         50  1999-07-20 MCA Universal Home Video
## 11         8.9         93  2004-03-09       Universal Pictures
## 12         6.5         59  2000-10-10       Universal Pictures
## 13         8.6         90  1999-11-02       Paramount Pictures
## 14         7.1         65  2002-03-05               Dreamworks
## 15         8.1         76  2003-05-06           DreamWorks SKG
## 16         7.3         55  2004-11-23           DreamWorks SKG
## 17         6.2         65  2008-10-14       Paramount Pictures
## 18         6.5         73  2005-11-22       Paramount Pictures
##                                                              actors
## 1             Sam Neill, Laura Dern, Jeff Goldblum, Richard Attenborough
## 2  Daniel Day-Lewis, Sally Field, David Strathairn, Joseph Gordon-Levitt
## 3                     Jamie Bell, Andy Serkis, Daniel Craig, Nick Frost
```

```
## 4              Jeremy Irvine, Peter Mullan, Emily Watson, Niels Arestrup
## 5     Mark Rylance, Domenick Lombardozzi, Victor Verhaeghe, Mark Fichera
## 6          Mark Rylance, Ruby Barnhill, Penelope Wilton, Jemaine Clement
## 7                  Dan Aykroyd, Ned Beatty, John Belushi, Lorraine Gary
## 8                Harrison Ford, Karen Allen, Paul Freeman, Ronald Lacey
## 9            Dee Wallace, Henry Thomas, Peter Coyote, Robert MacNaughton
## 10            Richard Dreyfuss, Holly Hunter, Brad Johnson, John Goodman
## 11            Liam Neeson, Ben Kingsley, Ralph Fiennes, Caroline Goodall
## 12      Jeff Goldblum, Julianne Moore, Pete Postlethwaite, Arliss Howard
## 13                  Tom Hanks, Tom Sizemore, Edward Burns, Barry Pepper
## 14         Haley Joel Osment, Frances O'Connor, Sam Robards, Jake Thomas
## 15        Leonardo DiCaprio, Tom Hanks, Christopher Walken, Martin Sheen
## 16           Tom Hanks, Catherine Zeta-Jones, Stanley Tucci, Chi McBride
## 17              Harrison Ford, Cate Blanchett, Karen Allen, Shia LaBeouf
## 18              Tom Cruise, Dakota Fanning, Miranda Otto, Justin Chatwin
##    imdb_votes         director release_date runtime
## 1      665919 Steven Spielberg   1993-06-11 127 min
## 2      207962 Steven Spielberg   2012-11-16 150 min
## 3      185185 Steven Spielberg   2011-12-21 107 min
## 4      122940 Steven Spielberg   2011-12-25 146 min
## 5      218441 Steven Spielberg   2015-10-16 142 min
## 6       51164 Steven Spielberg   2016-07-01 117 min
## 7       26731 Steven Spielberg   1979-12-14 118 min
## 8      718669 Steven Spielberg   1981-06-12 115 min
## 9      303850 Steven Spielberg   1982-06-11 115 min
## 10      23710 Steven Spielberg   1989-12-22 122 min
## 11     952369 Steven Spielberg   1994-02-04 195 min
## 12     302620 Steven Spielberg   1997-05-23 129 min
## 13     968995 Steven Spielberg   1998-07-24 169 min
## 14     254880 Steven Spielberg   2001-06-29 146 min
## 15     605176 Steven Spielberg   2002-12-25 141 min
## 16     328750 Steven Spielberg   2004-06-18 128 min
## 17     356423 Steven Spielberg   2008-05-22 122 min
## 18     356040 Steven Spielberg   2005-06-29 116 min
##                          genre
## 1   Adventure, Sci-Fi, Thriller
## 2     Biography, Drama, History
## 3  Animation, Action, Adventure
## 4                   Drama, War
## 5      Drama, History, Thriller
## 6    Adventure, Family, Fantasy
## 7           Action, Comedy, War
## 8             Action, Adventure
## 9                Family, Sci-Fi
## 10             Fantasy, Romance
## 11    Biography, Drama, History
## 12    Action, Adventure, Sci-Fi
## 13                  Drama, War
## 14     Adventure, Drama, Sci-Fi
## 15      Biography, Crime, Drama
## 16       Comedy, Drama, Romance
## 17   Action, Adventure, Fantasy
## 18  Adventure, Sci-Fi, Thriller
##                                                                  awards
```

```
## 1                     Won 3 Oscars. Another 28 wins & 17 nominations.
## 2                   Won 2 Oscars. Another 108 wins & 242 nominations.
## 3             Nominated for 1 Oscar. Another 22 wins & 60 nominations.
## 4            Nominated for 6 Oscars. Another 16 wins & 70 nominations.
## 5                      Won 1 Oscar. Another 30 wins & 97 nominations.
## 6                                          2 wins & 19 nominations.
## 7                      Nominated for 3 Oscars. Another 3 nominations.
## 8                     Won 4 Oscars. Another 30 wins & 23 nominations.
## 9                     Won 4 Oscars. Another 47 wins & 33 nominations.
## 10                                                    3 nominations.
## 11                    Won 7 Oscars. Another 72 wins & 37 nominations.
## 12             Nominated for 1 Oscar. Another 4 wins & 23 nominations.
## 13                    Won 5 Oscars. Another 74 wins & 74 nominations.
## 14           Nominated for 2 Oscars. Another 16 wins & 67 nominations.
## 15           Nominated for 2 Oscars. Another 13 wins & 41 nominations.
## 16                                           5 wins & 4 nominations.
## 17 Nominated for 1 BAFTA Film Award. Another 10 wins & 34 nominations.
## 18           Nominated for 3 Oscars. Another 14 wins & 44 nominations.
##      Budget Box.Office.Gross release.year release.month release.day
## 1  6.30e+07        402453882         1993             6          11
## 2  6.50e+07        182207973         2012            11          16
## 3  1.35e+08         77591831         2011            12          21
## 4  6.60e+07         79884879         2011            12          25
## 5  4.00e+07         72313754         2015            10          16
## 6  1.40e+08         55483770         2016             7           1
## 7  3.50e+07         31755742         1979            12          14
## 8  1.80e+07        248159971         1981             6          12
## 9  1.05e+07        435110554         1982             6          11
## 10 3.10e+07         43858790         1989            12          22
## 11 2.20e+07         96067179         1994             2           4
## 12 7.30e+07        229086679         1997             5          23
## 13 7.00e+07        216540909         1998             7          24
## 14 1.00e+08         78616689         2001             6          29
## 15 5.20e+07        164615351         2002            12          25
## 16 6.00e+07         77872883         2004             6          18
## 17 1.85e+08        317101119         2008             5          22
## 18 1.32e+08        234280354         2005             6          29
##            roi  n
## 1   5.38815686 18
## 2   1.80319958 18
## 3  -0.42524570 18
## 4   0.21037695 18
## 5   0.80784385 18
## 6  -0.60368736 18
## 7  -0.09269309 18
## 8  12.78666506 18
## 9  40.43910038 18
## 10  0.41479968 18
## 11  3.36668995 18
## 12  2.13817368 18
## 13  2.09344156 18
## 14 -0.21383311 18
## 15  2.16567983 18
## 16  0.29788138 18
```

```
## 17  0.71406010 18
## 18  0.77485117 18
```

For adaptions, I saw Alice in Wonderland, Beauty and Beast, batman V superman, the three musketers, pride & prejudice here.

Select movie franchise for review.

a. star war series, 3 here.

```
titles <- movies.before$title
star.war.index <- grep(pattern = "Star War", x=titles)
star.war <- movies.before[star.war.index,]
star.war %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                                     title release_date imdb_rating
## 1 Star Wars: Episode I - The Phantom Menace   1999-05-19         6.5
## 2                 Star Wars: The Clone Wars   2008-08-15         5.9
## 3              Star Wars: The Force Awakens   2015-12-18         8.1
##     Budget Box.Office.Gross      roi
## 1 1.15e+08        474544677 3.126475
## 2 8.50e+06         35161554 3.136653
## 3 2.45e+08        936662225 2.823111
```

b. Hobbit series, 3 here.

```
hobbit.index <- grep(pattern = "Hobbit", x=titles)
hobbit <- movies.before[hobbit.index,]
hobbit %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                                     title release_date imdb_rating
## 1         The Hobbit: An Unexpected Journey   2012-12-14         7.9
## 2       The Hobbit: The Desolation of Smaug   2013-12-13         7.9
## 3 The Hobbit: The Battle of the Five Armies   2014-12-17         7.4
##     Budget Box.Office.Gross        roi
## 1 1.80e+08        303003568 0.68335316
## 2 2.25e+08        258366855 0.14829713
## 3 2.50e+08        255119788 0.02047915
```

c. Fast and Furious, 7 here.

```
ff.index <- grep(pattern = "Furious", x=titles)
ff <- movies.before[ff.index,]
ff %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                                  title release_date imdb_rating  Budget
## 1              The Fast and the Furious   2001-06-22         6.7 3.8e+07
## 2                      2 Fast 2 Furious   2003-06-06         5.9 7.6e+07
## 3 The Fast and the Furious: Tokyo Drift   2006-06-16         6.0 8.5e+07
## 4                        Fast & Furious   2009-04-03         6.6 8.5e+07
## 5                      Fast & Furious 6   2013-05-24         7.1 1.6e+08
## 6                              Furious 7   2015-04-03         7.2 1.9e+08
## 7               The Fate of the Furious   2017-04-14         7.1 2.5e+08
```

```
##   Box.Office.Gross         roi
## 1        144533925  2.80352434
## 2        127154901  0.67309080
## 3         62514415 -0.26453629
## 4        155064265  0.82428547
## 5        238679850  0.49174906
## 6        353007020  0.85793168
## 7        225764765 -0.09694094
```

d. trnasformer series, 4 here.

```
trans.index <- grep(pattern = "Transformers", x=titles)
transformers <- movies.before[trans.index,]
transformers %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                               title release_date imdb_rating   Budget
## 1 Transformers: Revenge of the Fallen   2009-06-24         6.0 2.00e+08
## 2      Transformers: Dark of the Moon   2011-06-29         6.3 1.95e+08
## 3     Transformers: Age of Extinction   2014-06-27         5.7 2.10e+08
## 4       Transformers: The Last Knight   2017-06-21         5.5 2.17e+08
##   Box.Office.Gross        roi
## 1        402111870  1.0105594
## 2        352390543  0.8071310
## 3        245439076  0.1687575
## 4        130120862 -0.4003647
```

e. spiderman series, 5 here.

```
spider.index <- grep(pattern = "Spider-Man", x=titles)
spiderman <- movies.before[spider.index,]
spiderman %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                      title release_date imdb_rating   Budget
## 1             Spider-Man 2   2004-06-30         7.3 2.00e+08
## 2             Spider-Man 3   2007-05-04         6.2 2.58e+08
## 3   The Amazing Spider-Man   2012-07-03         7.0 2.30e+08
## 4 The Amazing Spider-Man 2   2014-05-02         6.7 2.00e+08
## 5   Spider-Man: Homecoming   2017-07-07         NaN 1.75e+08
##   Box.Office.Gross        roi
## 1        373585825 0.86792912
## 2        336530303 0.30438102
## 3        262030663 0.13926375
## 4        202853933 0.01426967
## 5        314057748 0.79461570
```

f. Harry Potter series, only 2 here.

```
hp.index <- grep(pattern = "Harry Potter", x=titles)
hp <- movies.before[hp.index,]
hp %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                                     title release_date imdb_rating
```

```
## 1       Harry Potter and the Half-Blood Prince   2009-07-15         7.5
## 2 Harry Potter and the Deathly Hallows: Part 2   2011-07-15         8.1
##     Budget Box.Office.Gross       roi
## 1 2.50e+08        301959197 0.2078368
## 2 1.25e+08        381011219 2.0480898
```

g. Jurassic Park, 3 here.

```
jp.index <- grep(pattern = "Jurassic", x=titles)
jp <- movies.before[jp.index,]
jp %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                           title release_date imdb_rating  Budget
## 1                 Jurassic Park   1993-06-11         8.1 6.3e+07
## 2 The Lost World: Jurassic Park   1997-05-23         6.5 7.3e+07
## 3               Jurassic World   2015-06-12         7.0 1.5e+08
##   Box.Office.Gross      roi
## 1        402453882 5.388157
## 2        229086679 2.138174
## 3        652270625 3.348471
```

h. Batman, 3 here.

```
bat.index <- grep(pattern = "Batman", x=titles)
batman <- movies.before[bat.index,]
batman %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                               title release_date imdb_rating  Budget
## 1 Batman v Superman: Dawn of Justice   2016-03-25         6.7 2.5e+08
## 2           Batman: The Killing Joke   2016-07-25         6.5 3.5e+06
## 3              The LEGO Batman Movie   2017-02-10         7.4 8.0e+07
##   Box.Office.Gross        roi
## 1        330360194 0.32144078
## 2          3775000 0.07857143
## 3        175750384 1.19687980
```

From our choice of 7 sequel movie seires, although we only have part of data in each series, it seems that only Star Wars and Jurassic Park has remained high roi for all series.

Adding all of these sequels into a dataframe:

```
sequels <- rbind(ff,hobbit,hp,jp,spiderman,star.war,transformers,batman)

sequels$category <- NA
sequels$category[1:7] <- "FF"
sequels$category[8:10] <- "Hobbit"
sequels$category[11:12] <- "HP"
sequels$category[13:15] <- "JP"
sequels$category[16:20] <- "Spiderman"
sequels$category[21:23] <- "SW"
sequels$category[24:27] <- "Transformers"
sequels$category[28:30] <- "Batman"

sequels
```

```
##         imdbid                                        title rating
## 8    tt0232500                    The Fast and the Furious  PG-13
## 585  tt1905041                            Fast & Furious 6  PG-13
## 866  tt2820852                                    Furious 7  PG-13
## 1133 tt4630562                     The Fate of the Furious  PG-13
## 1646 tt0322259                           2 Fast 2 Furious  PG-13
## 1831 tt0463985        The Fast and the Furious: Tokyo Drift  PG-13
## 2090 tt1013752                              Fast & Furious  PG-13
## 96   tt0903624           The Hobbit: An Unexpected Journey  PG-13
## 172  tt1170358         The Hobbit: The Desolation of Smaug  PG-13
## 740  tt2310332   The Hobbit: The Battle of the Five Armies  PG-13
## 185  tt1201607 Harry Potter and the Deathly Hallows: Part 2  PG-13
## 1755 tt0417741       Harry Potter and the Half-Blood Prince     PG
## 4    tt0107290                               Jurassic Park  PG-13
## 11   tt0369610                             Jurassic World  PG-13
## 1446 tt0119567               The Lost World: Jurassic Park  PG-13
## 105  tt0948470                      The Amazing Spider-Man  PG-13
## 575  tt1872181                    The Amazing Spider-Man 2  PG-13
## 720  tt2250912                      Spider-Man: Homecoming    N/A
## 1639 tt0316654                                 Spider-Man 2  PG-13
## 1748 tt0413300                                 Spider-Man 3  PG-13
## 7    tt0120915    Star Wars: Episode I - The Phantom Menace     PG
## 811  tt2488496                 Star Wars: The Force Awakens  PG-13
## 2235 tt1185834                    Star Wars: The Clone Wars     PG
## 314  tt1399103               Transformers: Dark of the Moon  PG-13
## 667  tt2109248              Transformers: Age of Extinction  PG-13
## 971  tt3371366               Transformers: The Last Knight  PG-13
## 2123 tt1055369          Transformers: Revenge of the Fallen  PG-13
## 905  tt2975590           Batman v Superman: Dawn of Justice  PG-13
## 1080 tt4116284                        The LEGO Batman Movie     PG
## 2594 tt4853102                     Batman: The Killing Joke      R
##      imdb_rating metacritic dvd_release                    production
## 8            6.7         58  2002-01-01             Universal Pictures
## 585          7.1         61  2013-12-10             Universal Pictures
## 866          7.2         67  2015-09-15             Universal Pictures
## 1133         7.1         56        <NA>             Universal Pictures
## 1646         5.9         38  2003-09-30 Universal Pictures Distributio
## 1831         6.0         45  2006-09-26             Universal Pictures
## 2090         6.6         46  2009-07-27             Universal Pictures
## 96           7.9         58  2013-03-19                  Warner Bros.
## 172          7.9         66  2014-04-08                  Warner Bros.
## 740          7.4         59  2015-03-24                  Warner Bros.
## 185          8.1         87  2011-11-11          Warner Bros. Pictures
## 1755         7.5         78  2009-12-08          Warner Bros. Pictures
## 4            8.1         68  2000-10-10         Universal City Studios
## 11           7.0         59  2015-10-20             Universal Pictures
## 1446         6.5         59  2000-10-10             Universal Pictures
## 105          7.0         66  2012-11-09                  Sony Pictures
## 575          6.7         53  2014-08-19                  Sony Pictures
## 720          NaN        NaN        <NA>                  Sony Pictures
## 1639         7.3         83  2004-11-30                  Sony Pictures
## 1748         6.2         59  2007-10-30                  Sony Pictures
## 7            6.5         51  2001-10-16               20th Century Fox
## 811          8.1         81  2016-04-05           Walt Disney Pictures
```

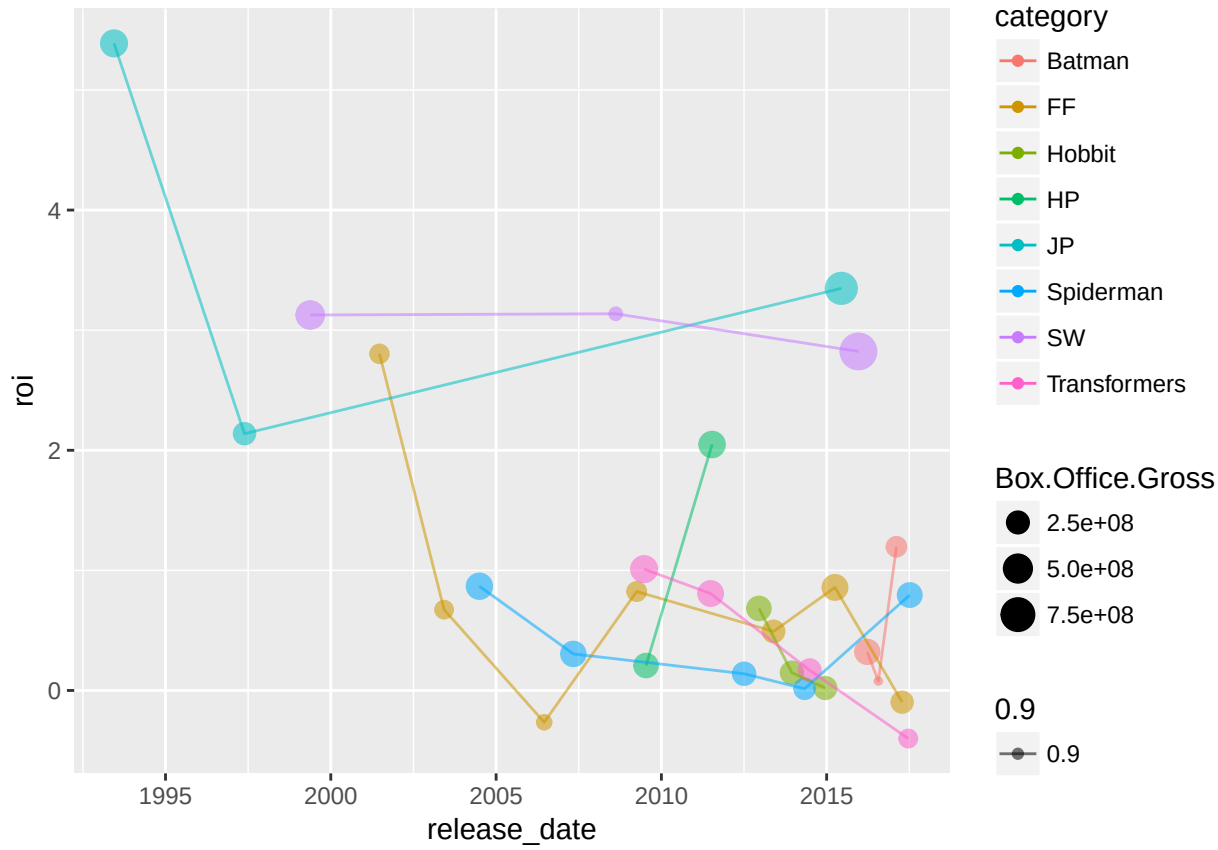
```
## 2235         5.9         35  2008-11-11          Warner Bros. Pictures
## 314          6.3         42  2011-09-30              Paramount Studios
## 667          5.7         32  2014-09-30             Paramount Pictures
## 971          5.5         30        <NA>             Paramount Pictures
## 2123         6.0         35  2009-10-20           Paramount/Dreamworks
## 905          6.7         44  2016-07-19          Warner Bros. Pictures
## 1080         7.4         75  2017-06-13          Warner Bros. Pictures
## 2594         6.5        NaN  2016-08-02              The Answer Studio
##                                                                  actors
## 8          Paul Walker, Vin Diesel, Michelle Rodriguez, Jordana Brewster
## 585            Vin Diesel, Paul Walker, Dwayne Johnson, Jordana Brewster
## 866           Vin Diesel, Paul Walker, Jason Statham, Michelle Rodriguez
## 1133       Vin Diesel, Jason Statham, Dwayne Johnson, Michelle Rodriguez
## 1646                Paul Walker, Tyrese Gibson, Eva Mendes, Cole Hauser
## 1831     Lucas Black, Damien Marzette, Trula M. Marcus, Zachery Ty Bryan
## 2090       Vin Diesel, Paul Walker, Jordana Brewster, Michelle Rodriguez
## 96             Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott
## 172            Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott
## 740            Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott
## 185        Ralph Fiennes, Michael Gambon, Alan Rickman, Daniel Radcliffe
## 1755      Daniel Radcliffe, Michael Gambon, Dave Legeno, Elarica Johnson
## 4            Sam Neill, Laura Dern, Jeff Goldblum, Richard Attenborough
## 11       Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio
## 1446     Jeff Goldblum, Julianne Moore, Pete Postlethwaite, Arliss Howard
## 105                Andrew Garfield, Emma Stone, Rhys Ifans, Denis Leary
## 575                Andrew Garfield, Emma Stone, Jamie Foxx, Dane DeHaan
## 720             Tom Holland, Chris Evans, Robert Downey Jr., Martin Starr
## 1639            Tobey Maguire, Kirsten Dunst, James Franco, Alfred Molina
## 1748     Tobey Maguire, Kirsten Dunst, James Franco, Thomas Haden Church
## 7                 Liam Neeson, Ewan McGregor, Natalie Portman, Jake Lloyd
## 811               Harrison Ford, Mark Hamill, Carrie Fisher, Adam Driver
## 2235 Matt Lanter, Ashley Eckstein, James Arnold Taylor, Dee Bradley Baker
## 314  Shia LaBeouf, Rosie Huntington-Whiteley, Josh Duhamel, John Turturro
## 667           Mark Wahlberg, Stanley Tucci, Kelsey Grammer, Nicola Peltz
## 971          Mark Wahlberg, Anthony Hopkins, Josh Duhamel, Laura Haddock
## 2123               Shia LaBeouf, Megan Fox, Josh Duhamel, Tyrese Gibson
## 905               Ben Affleck, Henry Cavill, Amy Adams, Jesse Eisenberg
## 1080            Will Arnett, Michael Cera, Rosario Dawson, Ralph Fiennes
## 2594                    Kevin Conroy, Mark Hamill, Tara Strong, Ray Wise
##      imdb_votes         director release_date runtime
## 8        290520        Rob Cohen   2001-06-22 106 min
## 585      318463       Justin Lin   2013-05-24 130 min
## 866      301880        James Wan   2015-04-03 137 min
## 1133      80260     F. Gary Gray   2017-04-14 136 min
## 1646     209662   John Singleton   2003-06-06 107 min
## 1831     198162       Justin Lin   2006-06-16 104 min
## 2090     222065       Justin Lin   2009-04-03 107 min
## 96       669541    Peter Jackson   2012-12-14 169 min
## 172      515177    Peter Jackson   2013-12-13 161 min
## 740      385820    Peter Jackson   2014-12-17 144 min
## 185      594362      David Yates   2011-07-15 130 min
## 1755     357718      David Yates   2009-07-15 153 min
## 4        665919 Steven Spielberg   1993-06-11 127 min
## 11       458076  Colin Trevorrow   2015-06-12 124 min
```

```
## 1446     302620 Steven Spielberg   1997-05-23 129 min
## 105      474291        Marc Webb   2012-07-03 136 min
## 575      344178        Marc Webb   2014-05-02 142 min
## 720         NaN        Jon Watts   2017-07-07     N/A
## 1639     448894        Sam Raimi   2004-06-30 127 min
## 1748     415052        Sam Raimi   2007-05-04 139 min
## 7        573249     George Lucas   1999-05-19 136 min
## 811      665521      J.J. Abrams   2015-12-18 136 min
## 2235      45411      Dave Filoni   2008-08-15  98 min
## 314      339695      Michael Bay   2011-06-29 154 min
## 667      256118      Michael Bay   2014-06-27 165 min
## 971         893      Michael Bay   2017-06-21 149 min
## 2123     342103      Michael Bay   2009-06-24 150 min
## 905      477874      Zack Snyder   2016-03-25 151 min
## 1080      57443      Chris McKay   2017-02-10 104 min
## 2594      36099          Sam Liu   2016-07-25  76 min
##                              genre
## 8          Action, Crime, Thriller
## 585        Action, Crime, Thriller
## 866        Action, Crime, Thriller
## 1133     Action, Adventure, Crime
## 1646       Action, Crime, Thriller
## 1831       Action, Crime, Thriller
## 2090       Action, Crime, Thriller
## 96              Adventure, Fantasy
## 172             Adventure, Fantasy
## 740             Adventure, Fantasy
## 185     Adventure, Drama, Fantasy
## 1755   Adventure, Family, Fantasy
## 4     Adventure, Sci-Fi, Thriller
## 11      Action, Adventure, Sci-Fi
## 1446    Action, Adventure, Sci-Fi
## 105              Action, Adventure
## 575     Action, Adventure, Sci-Fi
## 720     Action, Adventure, Sci-Fi
## 1639             Action, Adventure
## 1748             Action, Adventure
## 7      Action, Adventure, Fantasy
## 811    Action, Adventure, Fantasy
## 2235 Animation, Action, Adventure
## 314     Action, Adventure, Sci-Fi
## 667     Action, Adventure, Sci-Fi
## 971     Action, Adventure, Sci-Fi
## 2123    Action, Adventure, Sci-Fi
## 905     Action, Adventure, Sci-Fi
## 1080 Animation, Action, Adventure
## 2594     Animation, Action, Crime
##                                                                    awards
## 8                                                  10 wins & 12 nominations.
## 585                                                 8 wins & 21 nominations.
## 866     Nominated for 1 Golden Globe. Another 23 wins & 27 nominations.
## 1133                                                                  N/A
## 1646                                                4 wins & 13 nominations.
## 1831                                                  1 win & 4 nominations.
```

```
## 2090                                              5 wins & 2 nominations.
## 96            Nominated for 3 Oscars. Another 10 wins & 72 nominations.
## 172           Nominated for 3 Oscars. Another 16 wins & 86 nominations.
## 740             Nominated for 1 Oscar. Another 7 wins & 47 nominations.
## 185           Nominated for 3 Oscars. Another 45 wins & 92 nominations.
## 1755            Nominated for 1 Oscar. Another 8 wins & 35 nominations.
## 4                      Won 3 Oscars. Another 28 wins & 17 nominations.
## 11                                             6 wins & 54 nominations.
## 1446            Nominated for 1 Oscar. Another 4 wins & 23 nominations.
## 105                                            2 wins & 31 nominations.
## 575                                            3 wins & 29 nominations.
## 720                                                               N/A
## 1639                     Won 1 Oscar. Another 23 wins & 59 nominations.
## 1748 Nominated for 1 BAFTA Film Award. Another 3 wins & 32 nominations.
## 7             Nominated for 3 Oscars. Another 25 wins & 60 nominations.
## 811          Nominated for 5 Oscars. Another 51 wins & 115 nominations.
## 2235                                                        3 nominations.
## 314           Nominated for 3 Oscars. Another 10 wins & 39 nominations.
## 667                                            5 wins & 23 nominations.
## 971                                                               N/A
## 2123           Nominated for 1 Oscar. Another 15 wins & 27 nominations.
## 905                                           12 wins & 29 nominations.
## 1080                                           2 wins & 2 nominations.
## 2594                                            1 win & 2 nominations.
##        Budget Box.Office.Gross release.year release.month release.day
## 8    3.80e+07        144533925         2001             6           22
## 585  1.60e+08        238679850         2013             5           24
## 866  1.90e+08        353007020         2015             4            3
## 1133 2.50e+08        225764765         2017             4           14
## 1646 7.60e+07        127154901         2003             6            6
## 1831 8.50e+07         62514415         2006             6           16
## 2090 8.50e+07        155064265         2009             4            3
## 96   1.80e+08        303003568         2012            12           14
## 172  2.25e+08        258366855         2013            12           13
## 740  2.50e+08        255119788         2014            12           17
## 185  1.25e+08        381011219         2011             7           15
## 1755 2.50e+08        301959197         2009             7           15
## 4    6.30e+07        402453882         1993             6           11
## 11   1.50e+08        652270625         2015             6           12
## 1446 7.30e+07        229086679         1997             5           23
## 105  2.30e+08        262030663         2012             7            3
## 575  2.00e+08        202853933         2014             5            2
## 720  1.75e+08        314057748         2017             7            7
## 1639 2.00e+08        373585825         2004             6           30
## 1748 2.58e+08        336530303         2007             5            4
## 7    1.15e+08        474544677         1999             5           19
## 811  2.45e+08        936662225         2015            12           18
## 2235 8.50e+06         35161554         2008             8           15
## 314  1.95e+08        352390543         2011             6           29
## 667  2.10e+08        245439076         2014             6           27
## 971  2.17e+08        130120862         2017             6           21
## 2123 2.00e+08        402111870         2009             6           24
## 905  2.50e+08        330360194         2016             3           25
## 1080 8.00e+07        175750384         2017             2           10
```

```
## 2594 3.50e+06          3775000         2016             7          25
##               roi     category
## 8     2.80352434           FF
## 585   0.49174906           FF
## 866   0.85793168           FF
## 1133 -0.09694094           FF
## 1646  0.67309080           FF
## 1831 -0.26453629           FF
## 2090  0.82428547           FF
## 96    0.68335316       Hobbit
## 172   0.14829713       Hobbit
## 740   0.02047915       Hobbit
## 185   2.04808975           HP
## 1755  0.20783679           HP
## 4     5.38815686           JP
## 11    3.34847083           JP
## 1446  2.13817368           JP
## 105   0.13926375    Spiderman
## 575   0.01426967    Spiderman
## 720   0.79461570    Spiderman
## 1639  0.86792912    Spiderman
## 1748  0.30438102    Spiderman
## 7     3.12647545           SW
## 811   2.82311112           SW
## 2235  3.13665341           SW
## 314   0.80713099 Transformers
## 667   0.16875750 Transformers
## 971  -0.40036469 Transformers
## 2123  1.01055935 Transformers
## 905   0.32144078       Batman
## 1080  1.19687980       Batman
## 2594  0.07857143       Batman
```

```
#color <- c("#999999","#000000","#E69F00","#56B4E9","#009E73","#FF6600","#0072B2")

sequels %>%
  ggplot(aes(x=release_date,y=roi,col=category,alpha=0.9))+
  geom_point(aes(size=Box.Office.Gross))+
  geom_line()
```



For adaptions and remakes, there're 2 adapted from fairy tales, and 2 from classic literature.

a. Alice in Wonderland

```
alice.index <- grep(pattern = "Alice in Wonderland", x=titles)
alice <- movies.before[alice.index,]
alice %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                   title release_date imdb_rating Budget Box.Office.Gross
## 120 Alice in Wonderland   2010-03-05         6.5  2e+08        334191110
##           roi
## 120 0.6709555
```

b. Beauty and Beast

```
bb.index <- grep(pattern = "Beauty and the Beast", x=titles)
bb <- movies.before[bb.index,]
bb %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                    title release_date imdb_rating  Budget Box.Office.Gross
## 3   Beauty and the Beast   1991-11-22         8.0 2.5e+07        218967620
## 860 Beauty and the Beast   2017-03-17         7.6 1.6e+08        504014165
##           roi
## 3   7.758705
## 860 2.150089
```

c. the Three Musketers

```
musketeers.index <- grep(pattern = "Musketeers", x=titles)
musketeers <- movies.before[musketeers.index,]
musketeers %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##                    title release_date imdb_rating  Budget Box.Office.Gross
## 379 The Three Musketeers   2011-10-21         5.8 7.5e+07         20377913
##            roi
## 379 -0.7282945
```

d. Pride & Prejudice

```
pp.index <- grep(pattern = "Pride & Prejudice", x=titles)
pp <- movies.before[pp.index,]
pp %>%
  arrange(release_date) %>%
  select(title, release_date, imdb_rating, Budget, Box.Office.Gross, roi)
```

```
##               title release_date imdb_rating  Budget Box.Office.Gross
## 1 Pride & Prejudice   2005-11-23         7.8 2.8e+07         38405088
##         roi
## 1 0.3716103
```

For adaptions and remakes, I think the box office and roi may be related to the popularity of the content and ratings. For some literary work that's less well-known and may not be well produced, it's natural that the roi is negative. So choosing the right Intellectual Property "IP" here is very important for remakes and adaptions, see Beauty and the Beast.

7. Actors Analysis

```
movies.before %>%
  count(actors,sort=TRUE) %>%
  top_n(10)
```

```
## # A tibble: 14 x 2
##    actors                                                          n
##    <fct>                                                       <int>
##  1 N/A                                                             7
##  2 Bradley Cooper, Ed Helms, Zach Galifianakis, Justin Bartha      3
##  3 Ian McKellen, Martin Freeman, Richard Armitage, Ken Stott       3
##  4 Adam Sandler, Andy Samberg, Selena Gomez, Kevin James           2
##  5 Adam Sandler, Kevin James, Chris Rock, David Spade              2
##  6 Ben Stiller, Chris Rock, David Schwimmer, Jada Pinkett Smith    2
##  7 Brendan Fraser, Rachel Weisz, John Hannah, Arnold Vosloo        2
##  8 Chris Hemsworth, Natalie Portman, Tom Hiddleston, Anthony Hopkins  2
##  9 Chris Pratt, Zoe Saldana, Dave Bautista, Vin Diesel             2
## 10 Jennifer Lawrence, Josh Hutcherson, Liam Hemsworth, Woody Harrel~  2
## 11 Johnny Depp, Mia Wasikowska, Helena Bonham Carter, Anne Hathaway   2
## 12 Patrick Wilson, Rose Byrne, Ty Simpkins, Lin Shaye              2
## 13 Prabhas, Rana Daggubati, Anushka Shetty, Tamannaah Bhatia       2
## 14 Sylvester Stallone, Jason Statham, Jet Li, Dolph Lundgren       2
```

Need to break down actor column into document term matrix to better analyze the relationship of actors and box office.

```
library(tm)

# or first name and last name will be seperated into two items
```

```
actor <- as.character(movies.before$actors)
actor <- gsub(",", "   ", actor, fixed = TRUE)
actor <- gsub(" ", ".", actor, fixed = TRUE)
actor <- gsub("...", " ", actor, fixed = TRUE)
actor <- gsub(".", "", actor, fixed = TRUE)

actor.source <- VectorSource(actor)
actor.corpus <- VCorpus(actor.source)
actor.corpus.clean <- tm_map(actor.corpus,removePunctuation)

actor.dtm <- DocumentTermMatrix(actor.corpus.clean)
actor.matrix <- as.matrix(actor.dtm)

term.freq <- colSums(actor.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
head(term.freq,50)
```

```
##         robertdeniro             tomhanks          adamsandler
##                   23                   22                   21
##           liamneeson          stevecarell        dwaynejohnson
##                   20                   19                   18
##         jasonbateman         markwahlberg           naomiwatts
##                   18                   18                   18
##       robertdowneyjr          brucewillis        juliannemoore
##                   18                   17                   17
##            mattdamon   matthewmcconaughey           owenwilson
##                   17                   17                   17
##       woodyharrelson              amyadams      jenniferaniston
##                   17                   16                   16
##           johnnydepp arnoldschwarzenegger         jasonstatham
##                   16                   15                   15
##          jeffbridges          markruffalo          nicolascage
##                   15                   15                   15
##        susansarandon          alecbaldwin        bradleycooper
##                   15                   14                   14
##        channingtatum           danaykroyd          dennisquaid
##                   14                   14                   14
##           emilyblunt         ewanmcgregor        georgeclooney
##                   14                   14                   14
##          jamesfranco          jamesmcavoy         nicolekidman
##                   14                   14                   14
##         russellcrowe          willferrell             zacefron
##                   14                   14                   14
##         aaroneckhart         annakendrick           benstiller
##                   13                   13                   13
##        brendanfraser          camerondiaz      catherinekeener
##                   13                   13                   13
##           chrisevans            jackblack            jimcarrey
##                   13                   13                   13
##         juliaroberts         kevincostner
##                   13                   13
```

The Top 50 actors who appear in our dataset by number is Robert Deniro, Tom Hanks, Adam Sandler, Liam Neeson, Steve Carell and so on. There're lots of big stars in this list including Robert Downey Jr,

Julianne Moore, Matt Damon, Amy Adams, Jennifer Aniston, Johnny Depp, Arnold Schwarzenegger, Matthew McConaughey.

Due to the number of predictors is larger than the number of observations, we only include a limited number of actors here (keep the well-known ones to see if they have a direct association with box office) based on how many movies they played in in this dataset (>=10).

```
actor.df.whole <- as.data.frame(actor.matrix)
```

```
actor.df.selected <- actor.df.whole[,which(colSums(actor.df.whole)>=10)]
dim(actor.df.selected)
```

```
## [1] 2614  131
```

```
actor.df.selected$box.office <- movies.before$Box.Office.Gross
actor.lm <- lm(box.office~.,data=actor.df.selected)
summary(actor.lm)
```

```
##
## Call:
## lm(formula = box.office ~ ., data = actor.df.selected)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -202848306  -30052692  -18188670   12062006  769860221
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           30145745    1715412  17.573  < 2e-16 ***
## aaroneckhart          24482735   19799383   1.237 0.216375
## adamsandler           55636185   16732738   3.325 0.000897 ***
## alecbaldwin            6600281   19068486   0.346 0.729270
## alpacino              -9103502   20693246  -0.440 0.660029
## amyadams              38503562   18724037   2.056 0.039851 *
## angelinajolie         68050363   23590057   2.885 0.003952 **
## annakendrick          22589208   19788364   1.142 0.253755
## annettebening         -9052484   20914544  -0.433 0.665174
## anthonyhopkins         6265197   23377974   0.268 0.788725
## arnoldschwarzenegger  21646590   18211853   1.189 0.234711
## benaffleck            31107332   21846593   1.424 0.154602
## benkingsley           23118769   21889574   1.056 0.291000
## benstiller            71988585   20592446   3.496 0.000481 ***
## billmurray            22657254   22521108   1.006 0.314492
## billpaxton            33401767   22867407   1.461 0.144232
## billybobthornton     -14576189   22585166  -0.645 0.518736
## bradleycooper         45795972   20220994   2.265 0.023612 *
## bradpitt              28259276   21122084   1.338 0.181051
## brendanfraser         11455329   19726591   0.581 0.561492
## brendangleeson       -13185152   22966034  -0.574 0.565942
## brucewillis            9146458   17485771   0.523 0.600965
## camerondiaz           45631504   20019944   2.279 0.022734 *
## cateblanchett         38074803   20812552   1.829 0.067457 .
## catherinekeener       -9848902   20454047  -0.482 0.630194
## channingtatum         14741371   19437571   0.758 0.448286
## charlizetheron         8786342   22031874   0.399 0.690074
## chiwetelejiofor       29631248   22871924   1.296 0.195258
```

```
## chrisevans            90436457   21066472   4.293 1.83e-05 ***
## chrishemsworth        70517744   23105263   3.052 0.002297 **
## chrispine             95603006   21357788   4.476 7.94e-06 ***
## colinfarrell         -13152399   20525362  -0.641 0.521720
## danaykroyd            13431249   18901142   0.711 0.477396
## dennisquaid          -12692133   19051475  -0.666 0.505344
## denzelwashington      25798925   20860748   1.237 0.216308
## dwaynejohnson         46894551   16847439   2.783 0.005419 **
## eddiemurphy           66506340   21451216   3.100 0.001955 **
## emilyblunt            -2143885   19340834  -0.111 0.911746
## emmastone             56850625   22008899   2.583 0.009849 **
## ewanmcgregor            -41757   19483601  -0.002 0.998290
## forestwhitaker        -2522878   21647916  -0.117 0.907233
## garyoldman            23539386   22096329   1.065 0.286840
## georgeclooney         -3810440   19619076  -0.194 0.846019
## gerardbutler          35635339   20788402   1.714 0.086618 .
## harrisonford         136656259   22510902   6.071 1.47e-09 ***
## harveykeitel         -24553817   22776822  -1.078 0.281131
## hughjackman           68799839   22662640   3.036 0.002424 **
## jackblack             19811230   20496825   0.967 0.333863
## jakegyllenhaal         2750067   23058880   0.119 0.905077
## jamesfranco           44028054   19734217   2.231 0.025767 *
## jamesmcavoy           10480830   19674539   0.533 0.594282
## jamiefoxx             17500501   23475275   0.745 0.456048
## jasonbateman          39636107   17373914   2.281 0.022612 *
## jasonstatham          -2588705   18554128  -0.140 0.889049
## jeffbridges           18098454   18787055   0.963 0.335467
## jenniferaniston       17727010   18224357   0.973 0.330792
## jennifergarner        -5976033   21961093  -0.272 0.785553
## jenniferlawrence     127082894   21187302   5.998 2.29e-09 ***
## jeremyrenner          24291682   21085325   1.152 0.249405
## jesseeisenberg        11982152   23346674   0.513 0.607838
## jimcarrey             58372261   19992941   2.920 0.003536 **
## joaquinphoenix       -14681461   21420055  -0.685 0.493151
## joeledgerton         -18078275   21564307  -0.838 0.401919
## johncreilly            9768512   20685546   0.472 0.636798
## johncusack             -817688   21816310  -0.037 0.970105
## johngoodman           39394249   20681788   1.905 0.056924 .
## johnnydepp            41553380   17770693   2.338 0.019450 *
## jonahhill             29840480   23702363   1.259 0.208161
## josephgordonlevitt    57830866   20981489   2.756 0.005889 **
## joshbrolin            -3775464   22176905  -0.170 0.864833
## juliannemoore        -15399451   17746628  -0.868 0.385621
## juliaroberts          16186206   19975090   0.810 0.417834
## justinlong            19061544   22148018   0.861 0.389518
## katehudson            -1436124   22749565  -0.063 0.949670
## katewinslet           33503545   20858355   1.606 0.108348
## katherineheigl        -5192381   23086976  -0.225 0.822072
## keanureeves            2236708   20312517   0.110 0.912327
## kevincostner           5037849   19842125   0.254 0.799596
## kevinhart             59875724   21190102   2.826 0.004756 **
## kevinjames            33973291   22060662   1.540 0.123689
## kevinspacey           12028141   20752955   0.580 0.562246
## kirstendunst          64482520   22885559   2.818 0.004877 **
```

```
## kristenstewart        88381730   22675861   3.898 9.97e-05 ***
## kristenwiig           48065406   21780283   2.207 0.027417 *
## lesliemann             4208026   23706808   0.178 0.859128
## liamneeson            34698371   16273588   2.132 0.033089 *
## mariabello           -24251021   22794074  -1.064 0.287469
## markruffalo           30935152   19205311   1.611 0.107359
## markwahlberg          45209168   17012666   2.657 0.007925 **
## mattdamon             52638576   17447265   3.017 0.002579 **
## matthewmcconaughey    29350768   17612386   1.666 0.095743 .
## merylstreep            7834207   20090104   0.390 0.696604
## michaeldouglas        12443508   20794474   0.598 0.549625
## michaelfassbender      4657864   20435650   0.228 0.819721
## michaelshannon       -34215098   23213799  -1.474 0.140632
## michellemonaghan     -20229316   22068199  -0.917 0.359403
## milakunis             21317524   21028001   1.014 0.310792
## morganfreeman         16647632   19945121   0.835 0.403983
## naomiwatts              -90187   16682688  -0.005 0.995687
## natalieportman        42359640   21474577   1.973 0.048658 *
## nicolascage           -7708237   18285232  -0.422 0.673386
## nicolekidman         -22553856   19400951  -1.163 0.245139
## owenwilson            20088867   17736469   1.133 0.257479
## paulrudd              30355415   23309215   1.302 0.192937
## philipseymourhoffman -18199399   21930217  -0.830 0.406688
## rachelmcadams         15714430   22046623   0.713 0.476048
## ralphfiennes          79187125   20104341   3.939 8.41e-05 ***
## robertdeniro          10872377   15320101   0.710 0.477969
## robertdowneyjr       127084346   18096512   7.023 2.80e-12 ***
## robertduvall          -4575720   20704636  -0.221 0.825111
## robertredford          -638116   22816950  -0.028 0.977691
## ronperlman             9571812   22231265   0.431 0.666828
## rosariodawson        -14176103   20954084  -0.677 0.498766
## russellcrowe          29542240   19131097   1.544 0.122667
## ryanreynolds          39873936   20414567   1.953 0.050907 .
## samuelljackson        27289787   22833141   1.195 0.232130
## scarlettjohansson     43934272   20384090   2.155 0.031233 *
## seanpenn              -3358833   23193414  -0.145 0.884866
## sethrogen              5250668   21076685   0.249 0.803287
## stevecarell           50841083   17077669   2.977 0.002939 **
## susansarandon        -14585021   18356336  -0.795 0.426951
## tomcruise             75886829   21959196   3.456 0.000558 ***
## tomhanks             102993538   15127933   6.808 1.24e-11 ***
## tommyleejones         15657368   20959214   0.747 0.455110
## tomwilkinson           1832594   19815013   0.092 0.926320
## vincevaughn            2622735   21842242   0.120 0.904432
## vindiesel            160996598   21511537   7.484 9.94e-14 ***
## willarnett            72017928   23137153   3.113 0.001875 **
## willferrell           25521606   19395695   1.316 0.188350
## woodyharrelson        13229312   18005429   0.735 0.462567
## zacefron              12436672   19004122   0.654 0.512902
## zachgalifianakis      43334741   23710211   1.828 0.067717 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69660000 on 2482 degrees of freedom
```

```
## Multiple R-squared:  0.2113, Adjusted R-squared:  0.1697
## F-statistic: 5.077 on 131 and 2482 DF,  p-value: < 2.2e-16
```

There're some actors with positive coefficients meaning positive impact on box office if they are in the movie, and some with negative coefficients. Some actors have p-value < 0.05, meaning they do make a difference, and more are not significant. We analyze the actors with significant influence here.

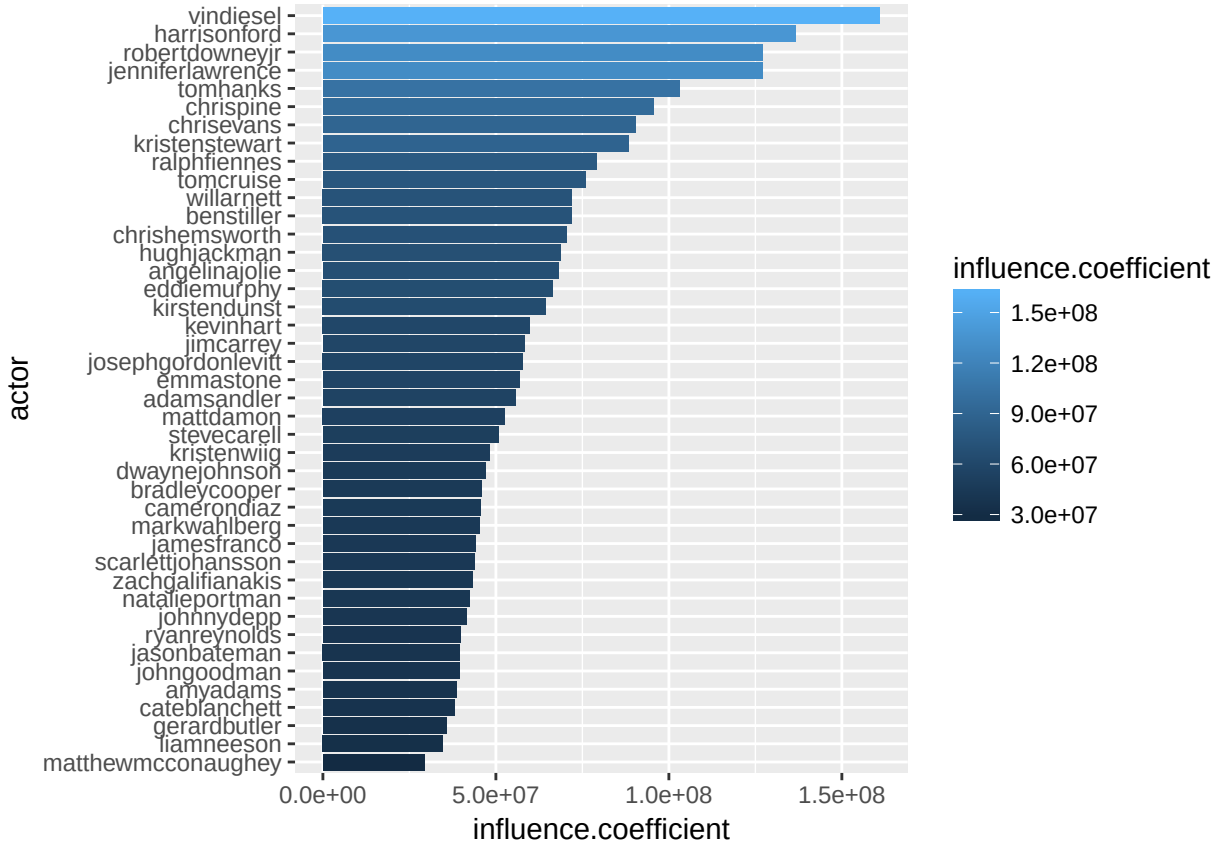
```
keep.actor <- c("zachgalifianakis","willarnett","vindiesel","tomhanks","tomcruise","stevecarell","scarl

actor.coeffi <- data.frame(actor=keep.actor, influence.coefficient=actor.lm$coefficients[keep.actor])
head(actor.coeffi)
```

```
##                              actor influence.coefficient
## zachgalifianakis zachgalifianakis              43334741
## willarnett             willarnett              72017928
## vindiesel               vindiesel             160996598
## tomhanks                 tomhanks             102993538
## tomcruise               tomcruise              75886829
## stevecarell           stevecarell              50841083
```

```
dim(actor.coeffi)
```

```
## [1] 42  2
```

```
actor.coeffi %>%
  ggplot(aes(x=reorder(actor,influence.coefficient),y=influence.coefficient,fill=influence.coefficient))
  geom_col()+
  coord_flip()+
  labs(x="actor")
```



All actors that have significant influence over box office have positive coefficients here. The top 5 actors who could bring considerable amount of box office are Vin Diesel, Harrison Ford, Robert Downey Jr, Jennifer Lawrence, and Tom Hanks. Generally speaking more actors than actresses make it to this graph, and this may have something to do with Hollywood's unfair payment between gender.

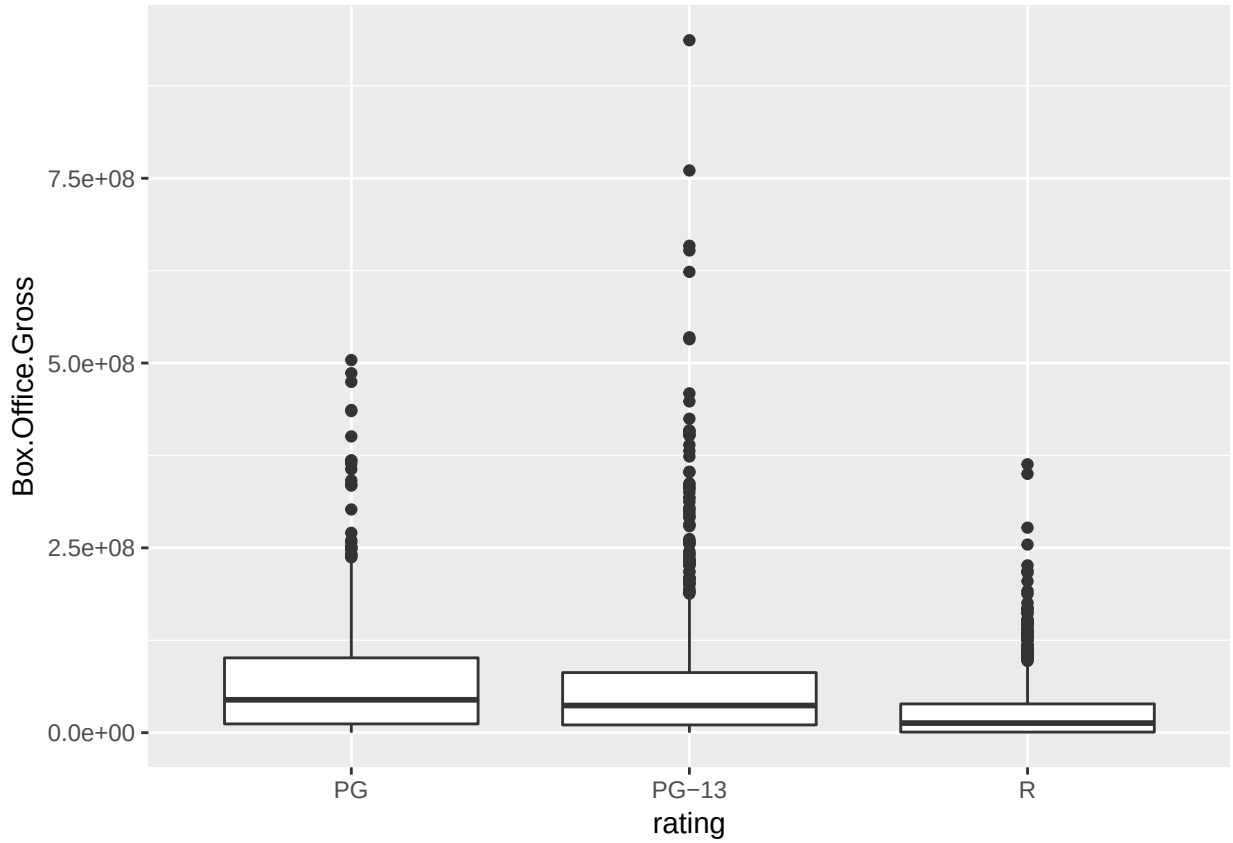
8. Does a movie's rating influence its box office?

```
table(movies.before$rating)
```

```
##
##             APPROVED            G            M          N/A       NC-17 NOT RATED
##            0        2           47            0          142          3           88
##           NR       PG        PG-13            R       TV-14        TV-G        TV-MA
##            1      345          888         1071            0           0            0
##        TV-PG     TV-Y     Unrated     UNRATED            X
##            1        0            0          25            1
```

The three most common ratings are R, PG-13, PG here, so we'll only analyze the difference among these three.

```
movies.before %>%
  filter(rating %in% c("R","PG","PG-13")) %>%
  group_by(rating) %>%
  ggplot(aes(x=rating,y=Box.Office.Gross))+
  geom_boxplot()
```



R rating movies mostly have lower box office than PG-13 and PG. It's easy to understand since higher restriction level limits the audience.

9. Why do some small budge films end up being blockbuster hits? Conversely, why do some large budge films fail?

We filter out movies that have top 10 roi and bottom 10 roi to find a pattern.

```
movies.before %>%
  arrange(desc(roi)) %>%
  top_n(10)
```

```
##       imdbid                    title rating imdb_rating metacritic
## 1  tt1179904      Paranormal Activity      R         6.3         68
## 2  tt3320578                   Veeram    N/A         6.7        NaN
## 3  tt5481184           League of Gods  PG-13         4.6        NaN
## 4  tt2309260              The Gallows      R         4.2         30
## 5  tt0077651                Halloween      R         7.8         78
## 6  tt0080761          Friday the 13th      R         6.5         19
## 7  tt1560985         The Devil Inside      R         4.2         18
## 8  tt1129423                Fireproof     PG         6.6         28
## 9  tt0259446 My Big Fat Greek Wedding     PG         6.6         62
## 10 tt0088847       The Breakfast Club      R         7.9         62
##    dvd_release                      production
## 1   2009-12-29             Paramount Pictures
## 2         <NA>               AIM Distribution
## 3         <NA>                            N/A
## 4   2015-10-13                New Line Cinema
## 5   1997-10-27 Compass International Pictures
## 6   1999-10-19             Paramount Pictures
## 7   2012-05-15             Paramount Pictures
## 8   2009-01-27              Sherwood Pictures
## 9   2003-02-11                      IFC Films
## 10  2003-09-02             Universal Pictures
##                                                                  actors
## 1           Katie Featherston, Micah Sloat, Mark Fredrichs, Amber Armstrong
## 2                            Ajith Kumar, Tamannaah Bhatia, Vidharth, Bala
## 3                          Jet Li, Bingbing Fan, Zhang Wen, Xiaoming Huang
## 4                 Reese Mishler, Pfeifer Brown, Ryan Shoos, Cassidy Gifford
## 5                Donald Pleasence, Jamie Lee Curtis, Nancy Kyes, P.J. Soles
## 6                Betsy Palmer, Adrienne King, Jeannine Taylor, Robbi Morgan
## 7              Fernanda Andrade, Simon Quarterman, Evan Helmuth, Ionut Grama
## 8                       Kirk Cameron, Erin Bethea, Ken Bevel, Stephen Dervan
## 9  Nia Vardalos, Michael Constantine, Christina Eleusiniotis, Kaylee Vieira
## 10         Emilio Estevez, Paul Gleason, Anthony Michael Hall, John Kapelos
##    imdb_votes                    director release_date runtime
## 1      194719                  Oren Peli   2009-10-16  86 min
## 2        5585                       Siva   2014-01-10 161 min
## 3         888                   Koan Hui   2016-07-29 109 min
## 4       15549 Travis Cluff, Chris Lofing   2015-07-10  81 min
## 5      171144             John Carpenter   1978-10-25  91 min
## 6       91250         Sean S. Cunningham   1980-05-09  95 min
## 7       31468         William Brent Bell   2012-01-06  83 min
## 8       18194              Alex Kendrick   2008-09-26 122 min
## 9      107647                 Joel Zwick   2002-08-02  95 min
## 10     281569                John Hughes   1985-02-15  97 min
##                         genre
## 1  Horror, Mystery, Thriller
```

```
## 2             Action, Family
## 3            Action, Fantasy
## 4           Horror, Thriller
## 5           Horror, Thriller
## 6  Horror, Mystery, Thriller
## 7                    Horror
## 8             Drama, Romance
## 9     Comedy, Drama, Romance
## 10             Comedy, Drama
##                                                    awards  Budget
## 1                                 3 wins & 12 nominations.   15000
## 2                                           2 nominations.      44
## 3                                            1 nomination.     300
## 4                                            1 nomination.  100000
## 5                                 5 wins & 2 nominations.  300000
## 6                                           5 nominations.  550000
## 7                                 2 wins & 3 nominations.  750000
## 8                                                     N/A  500000
## 9  Nominated for 1 Oscar. Another 20 wins & 28 nominations. 5000000
## 10                                                2 wins. 1000000
##    Box.Office.Gross release.year release.month release.day       roi
## 1         107918810         2009            10          16 7193.58733
## 2            243955         2014             1          10 5543.43182
## 3            143430         2016             7          29  477.10000
## 4          22757819         2015             7          10  226.57819
## 5          47000000         1978            10          25  155.66667
## 6          39754601         1980             5           9   71.28109
## 7          53261944         2012             1           6   70.01593
## 8          33456317         2008             9          26   65.91263
## 9         241438208         2002             8           2   47.28764
## 10         45875171         1985             2          15   44.87517
```

By observation, it's noticed that for movies that have high roi, they mostly have low budget, not-so-well-known director and actors, but have some nominations and awards. And half of them are horror+thriller genre.

```
movies.before %>%
  arrange(roi) %>%
  top_n(-10)
```

```
##       imdbid                            title    rating imdb_rating
## 1  tt1606384                          My Way         R         7.8
## 2  tt4832640                          Sultan NOT RATED         7.1
## 3  tt2557256                          Iceman         R         4.8
## 4  tt2215077            Half of a Yellow Sun         R         6.2
## 5  tt1487931                          Khumba NOT RATED         5.8
## 6  tt0273689             It's All About Love         R         5.5
## 7  tt0485947                       Mr. Nobody         R         7.9
## 8  tt2275471                    A Perfect Man         R         5.2
## 9  tt1857913 The Sorcerer and the White Snake     PG-13         5.9
## 10 tt1410063              The Flowers of War         R         7.6
##    metacritic dvd_release                production
## 1          30  2012-07-23          CJ Entertainment
## 2         NaN        <NA>            Yash Raj Films
## 3          32  2014-11-11               Well Go USA
## 4          51  2014-07-29            Monterey Media
```

```
## 5          40  2014-03-10   Millenium Entertainment
## 6          32  2005-09-27            Focus Features
## 7          63  2014-02-25         Magnolia Pictures
## 8          31  2014-01-27                IFC Films
## 9          41  2013-04-09         Magnolia Pictures
## 10         46  2012-07-10 Wrekin Hill Entertainment
##                                                                  actors
## 1              Dong-gun Jang, Joe Odagiri, Bingbing Fan, In-kwon Kim
## 2           Marko Zaror, Salman Khan, Anushka Sharma, Ron Smoorenburg
## 3                  Donnie Yen, Baoqiang Wang, Shengyi Huang, Kang Yu
## 4     Thandie Newton, Chiwetel Ejiofor, Anika Noni Rose, Joseph Mawle
## 5         Jake T. Austin, Adrian Rhodes, Sam Riegel, Bryce Papenbrook
## 6          Joaquin Phoenix, Claire Danes, Sean Penn, Douglas Henshall
## 7              Jared Leto, Sarah Polley, Diane Kruger, Linh Dan Pham
## 8  Jeanne Tripplehorn, Liev Schreiber, Joelle Carter, Louise Fletcher
## 9                  Jet Li, Shengyi Huang, Raymond Lam, Charlene Choi
## 10                Christian Bale, Ni Ni, Xinyi Zhang, Tianyuan Huang
##    imdb_votes           director release_date runtime
## 1        7915        Je-kyu Kang   2012-04-20 137 min
## 2       27797    Ali Abbas Zafar   2016-07-06 170 min
## 3        2307    Wing-Cheong Law   2014-09-19 104 min
## 4        1349       Biyi Bandele   2014-05-16 111 min
## 5        4449 Anthony Silverston   2013-10-25  85 min
## 6        7336  Thomas Vinterberg   2003-01-10 104 min
## 7      172168   Jaco Van Dormael   2013-09-26 141 min
## 8        1620   Kees Van Oostrum   2013-10-31  95 min
## 9        7194     Siu-Tung Ching   2011-09-28 100 min
## 10      41515        Yimou Zhang   2011-12-16 146 min
##                           genre
## 1        Action, Drama, History
## 2         Action, Drama, Family
## 3       Action, Comedy, History
## 4                 Drama, Romance
## 5  Animation, Adventure, Family
## 6      Drama, Thriller, Romance
## 7       Drama, Fantasy, Romance
## 8                 Drama, Romance
## 9                Action, Fantasy
## 10      Drama, History, Romance
##                                                    awards   Budget
## 1                                    1 win & 1 nomination. 3.00e+10
## 2                                  4 wins & 9 nominations. 7.00e+08
## 3                                                  1 win. 2.00e+08
## 4                                         3 nominations. 1.27e+09
## 5                                         3 nominations. 2.00e+07
## 6                                                3 wins. 8.60e+07
## 7                                  11 wins & 5 nominations. 4.70e+07
## 8                                                  N/A 5.00e+06
## 9                                    1 win & 4 nominations. 2.00e+08
## 10 Nominated for 1 Golden Globe. Another 5 wins & 12 nominations. 9.40e+07
##    Box.Office.Gross release.year release.month release.day        roi
## 1             67330         2012             4          20 -0.9999978
## 2              6173         2016             7           6 -0.9999912
## 3              7679         2014             9          19 -0.9999616
```

```
## 4             53645         2014             5          16 -0.9999578
## 5              1029         2013            10          25 -0.9999486
## 6              6140         2003             1          10 -0.9999286
## 7              3600         2013             9          26 -0.9999234
## 8               388         2013            10          31 -0.9999224
## 9             18445         2011             9          28 -0.9999078
## 10             9213         2011            12          16 -0.9999020
```

For movies that have low roi, their budget is high and box office failed to make ends meet. Half of them are action movies and half are drama+romance. Action movie is usually more costly than other genre as we mentioned above.

PREDICTIVE MODELING

1. In order to do Linear Regression, we don't want useless columns including title, imdbid, dvd_release, release_date (keep seperate release info) and roi so we only keep the useful numeric columns as predictors.

```
keep <- c("imdb_rating","metacritic","imdb_votes","Budget","Box.Office.Gross","release.year","release.m
movies.clean <- movies.before[keep]
```

```
head(movies.clean)
```

```
##   imdb_rating metacritic imdb_votes  Budget Box.Office.Gross release.year
## 1         7.2        NaN      17801 2.2e+05            46585         1965
## 2         7.8         67     289586 3.0e+07        238632124         1984
## 3         8.0         95     332843 2.5e+07        218967620         1991
## 4         8.1         68     665919 6.3e+07        402453882         1993
## 5         8.8         82    1365937 5.5e+07        330252182         1994
## 6         7.7         74     847267 2.0e+08        658672302         1997
##   release.month
## 1             5
## 2             6
## 3            11
## 4             6
## 5             7
## 6            12
```

3. For runtime column we remove the "min" and transform it into numeric type.

```
for (i in 1:nrow(movies.before)){
  movies.clean$runtime[i] <- as.numeric(strsplit(as.character(movies.before$runtime),split=" ")[[i]][1]
}
```

4. For other categorical columns including rating, production, actors, director, genre, awards, keywords, we will use bag of words method to create a dtm and add these dummy variable to train.clean.

```
library(tm)
```

4.1 Genre

```
genre <- as.character(movies.before$genre)
head(genre,5)
```

```
## [1] "Drama, Mystery, Sci-Fi"      "Action, Adventure, Comedy"
## [3] "Animation, Family, Fantasy"  "Adventure, Sci-Fi, Thriller"
## [5] "Comedy, Drama, Romance"
```

```
genre.source <- VectorSource(genre)
genre.corpus <- VCorpus(genre.source)
print(genre.corpus[[1]][1])
```

```
## $content
## [1] "Drama, Mystery, Sci-Fi"
```

Create a clean_corpus function.

```
clean_corpus <- function(corpus){
  corpus <- tm_map(corpus,removePunctuation)
  return(corpus)
}
```

```
genre.corpus.clean <- clean_corpus(genre.corpus)
print(genre.corpus.clean[[1]][1])
```

```
## $content
## [1] "Drama Mystery SciFi"
```

```
genre.dtm <- DocumentTermMatrix(genre.corpus.clean)
genre.matrix <- as.matrix(genre.dtm)
dim(genre.matrix)
```

```
## [1] 2614   23
```

```
genre.matrix[1:5,1:23]
```

```
##     Terms
## Docs action adventure animation biography comedy crime documentary drama
##    1      0         0         0         0      0     0           0     1
##    2      1         1         0         0      1     0           0     0
##    3      0         0         1         0      0     0           0     0
##    4      0         1         0         0      0     0           0     0
##    5      0         0         0         0      1     0           0     1
##     Terms
## Docs family fantasy history horror music musical mystery news romance
##    1      0       0       0      0     0       0       1    0       0
##    2      0       0       0      0     0       0       0    0       0
##    3      1       1       0      0     0       0       0    0       0
##    4      0       0       0      0     0       0       0    0       0
##    5      0       0       0      0     0       0       0    0       1
##     Terms
## Docs scifi short sport thriller war western
##    1     1     0     0        0   0       0
##    2     0     0     0        0   0       0
##    3     0     0     0        0   0       0
##    4     1     0     0        1   0       0
##    5     0     0     0        0   0       0
```

```
term.freq <- colSums(genre.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
term.freq
```

```
##       drama      comedy      action   adventure     romance    thriller
##        1320         965         633         486         451         423
##       crime      horror     mystery     fantasy       scifi   biography
```

```
##         413         245         212         200         183         177
##      family   animation documentary       music     history       sport
##         169         142          98          80          75          61
##         war     musical     western        news       short
##          41          22          15           4           1
```

```
keep.genre <- c("action","adventure","comedy","drama","romance","fantasy","horror","thriller","mystery"
genre.df <- as.data.frame(genre.matrix)[keep.genre]
head(genre.df)
```

```
##   action adventure comedy drama romance fantasy horror thriller mystery
## 1      0         0      0     1       0       0      0        0       1
## 2      1         1      1     0       0       0      0        0       0
## 3      0         0      0     0       0       1      0        0       0
## 4      0         1      0     0       0       0      0        1       0
## 5      0         0      1     1       1       0      0        0       0
## 6      0         0      0     1       1       0      0        0       0
##   scifi crime
## 1     1     0
## 2     0     0
## 3     0     0
## 4     1     0
## 5     0     0
## 6     0     0
```

4.2 Actor

```
actor.df <- as.data.frame(actor.matrix)[keep.actor]
head(actor.df)
```

```
##   zachgalifianakis willarnett vindiesel tomhanks tomcruise stevecarell
## 1                0          0         0        0         0           0
## 2                0          0         0        0         0           0
## 3                0          0         0        0         0           0
## 4                0          0         0        0         0           0
## 5                0          0         0        1         0           0
## 6                0          0         0        0         0           0
##   scarlettjohansson ryanreynolds robertdowneyjr ralphfiennes
## 1                 0            0              0            0
## 2                 0            0              0            0
## 3                 0            0              0            0
## 4                 0            0              0            0
## 5                 0            0              0            0
## 6                 0            0              0            0
##   natalieportman mattdamon matthewmcconaughey markwahlberg kristenwiig
## 1              0         0                  0            0           0
## 2              0         0                  0            0           0
## 3              0         0                  0            0           0
## 4              0         0                  0            0           0
## 5              0         0                  0            0           0
## 6              0         0                  0            0           0
##   liamneeson kristenstewart kirstendunst kevinhart johnnydepp
## 1          0              0            0         0          0
## 2          0              0            0         0          0
## 3          0              0            0         0          0
## 4          0              0            0         0          0
```

```
## 5          0              0            0         0          0
## 6          0              0            0         0          0
##   josephgordonlevitt johngoodman jasonbateman jenniferlawrence jimcarrey
## 1                  0           0           0                 0         0
## 2                  0           0           0                 0         0
## 3                  0           0           0                 0         0
## 4                  0           0           0                 0         0
## 5                  0           0           0                 0         0
## 6                  0           0           0                 0         0
##   jamesfranco hughjackman harrisonford gerardbutler adamsandler amyadams
## 1           0           0            0            0           0        0
## 2           0           0            0            0           0        0
## 3           0           0            0            0           0        0
## 4           0           0            0            0           0        0
## 5           0           0            0            0           0        0
## 6           0           0            0            0           0        0
##   angelinajolie benstiller bradleycooper camerondiaz cateblanchett
## 1             0          0             0           0             0
## 2             0          0             0           0             0
## 3             0          0             0           0             0
## 4             0          0             0           0             0
## 5             0          0             0           0             0
## 6             0          0             0           0             0
##   chrisevans chrishemsworth chrispine dwaynejohnson eddiemurphy emmastone
## 1          0              0         0             0           0         0
## 2          0              0         0             0           0         0
## 3          0              0         0             0           0         0
## 4          0              0         0             0           0         0
## 5          0              0         0             0           0         0
## 6          0              0         0             0           0         0
```

4.3 Production

```
production <- as.character(movies.before$production)
production <- gsub(" ","",production,fixed=TRUE)
production.source <- VectorSource(production)
production.corpus <- VCorpus(production.source)

production.dtm <- DocumentTermMatrix(production.corpus)
production.matrix <- as.matrix(production.dtm)

term.freq <- colSums(production.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
head(term.freq,20)
```

```
##     universalpictures   warnerbros.pictures        20thcenturyfox
##                   263                   140                   134
##          sonypictures     paramountpictures                   n/a
##                   109                    96                    69
##    waltdisneypictures mcauniversalhomevideo         focusfeatures
##                    69                    64                    62
##        lionsgatefilms  sonypicturesclassics              ifcfilms
##                    51                    44                    41
##      magnoliapictures   theweinsteincompany             lionsgate
##                    41                    40                    38
```

```
##   summitentertainment      columbiapictures       relativitymedia
##                    35                    33                    31
##   roadsideattractions         openroadfilms
##                    30                    29
```

```
keep.production <- c("universalpictures","warnerbros.pictures","20thcenturyfox","sonypictures","paramou
production.df <- as.data.frame(production.matrix)[keep.production]
```

4.4 Director

```
director <- as.character(movies.before$director)
director <- gsub(" ","",director,fixe=TRUE)
director.source <- VectorSource(director)
director.corpus <- VCorpus(director.source)

director.dtm <- DocumentTermMatrix(director.corpus)
director.matrix <- as.matrix(director.dtm)

term.freq <- colSums(director.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
```

```
head(term.freq,70)
```

```
##    stevenspielberg          ronhoward   stevensoderbergh
##                 18                 13                 10
##        dennisdugan     robertzemeckis          spikelee
##                  9                  9                  9
##        oliverstone          peterberg        ridleyscott
##                  8                  8                  8
##             anglee      clinteastwood          jonm.chu
##                  7                  7                  7
##         michaelbay   paulw.s.anderson          robcohen
##                  7                  7                  7
##     rolandemmerich          shawnlevy          timburton
##                  7                  7                  7
##       toddphillips         tylerperry         dannyboyle
##                  7                  7                  6
##   davidgordongreen        ivanreitman          jameswan
##                  6                  6                  6
## joelcoen,ethancoen       malcolmd.lee         markwaters
##                  6                  6                  6
##     paulgreengrass          paulweitz       peterjackson
##                  6                  6                  6
##   richardlinklater          robreiner          samraimi
##                  6                  6                  6
##         woodyallen         zacksnyder       antoinefuqua
##                  6                  6                  5
##      barrylevinson         billcondon        bryansinger
##                  5                  5                  5
##   christophernolan         d.j.caruso         davidkoepp
##                  5                  5                  5
##        davidlynch ethancoen,joelcoen      garrymarshall
##                  5                  5                  5
##      goreverbinski         gusvansant       jasonreitman
##                  5                  5                  5
##          joewright      johncarpenter         jonfavreau
```

```
##                 5                 5                 5
##         juddapatow          justinlin     kathrynbigelow
##                 5                 5                 5
##         petersegal   quentintarantino      robertluketic
##                 5                 5                 5
##    robertschwentke     rogerdonaldson          sammendes
##                 5                 5                 5
##       sofiacoppola     taylorhackford         tomshadyac
##                 5                 5                 5
##          wescraven          adammckay       annefletcher
##                 5                 4                 4
##        brettratner       briandepalma     brianhelgeland
##                 4                 4                 4
##        brianlevant
##                 4
```

```
keep.director <- c("stevenspielberg","jamescameron","christophernolan","woodyallen","ronhoward","robertz
director.df <- as.data.frame(director.matrix)[keep.director]
```

4.5 Award and nominaton

Here the infos about awards and nomination has been reduced to four columns: nomination, oscar, golden globe, and award. Numbers of awards and nominations are not included for simplicity reason.

```
library(textreg)
```

```
award <- as.character(movies.before$awards)
award.source <- VectorSource(award)
award.corpus <- VCorpus(award.source)

# clean and stem corpus
award.corpus <- tm_map(award.corpus,removePunctuation)
award.corpus <- tm_map(award.corpus,removeNumbers)
award.corpus <- tm_map(award.corpus,content_transformer(tolower))
award.corpus <- stem.corpus(award.corpus,verbose=FALSE)

award.dtm <- DocumentTermMatrix(award.corpus)
award.matrix <- as.matrix(award.dtm)

term.freq <- colSums(award.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
term.freq
```

```
## nomin+   win+ anoth+ oscar+    for    won globe+ golden award+  bafta
##   2434   1640    516    382    379    137    108    108     26     26
##   film
##     26
```

```
keep.award <- c("nomin+","oscar+","globe+","award+")
award.df <- as.data.frame(award.matrix)[keep.award]
colnames(award.df) <- c("nomination","oscar","golden globe","award")
```

4.6 Keyword

```
# keyword <- as.character(movies.before$keywords) executed above
keyword <- gsub("|"," ",keyword,fixe=TRUE)
keyword <- gsub("-"," ",keyword,fixe=TRUE)
```

```r
keyword.source <- VectorSource(keyword)
keyword.corpus <- VCorpus(keyword.source)
keyword.corpus <- tm_map(keyword.corpus,removeWords,stopwords("en"))

keyword.dtm <- DocumentTermMatrix(keyword.corpus)
keyword.matrix <- as.matrix(keyword.dtm)

term.freq <- colSums(keyword.matrix)
term.freq <- sort(term.freq,decreasing = TRUE)
head(term.freq,100)
```

```
##    reference relationship        death         shot        title
##         5719         5611         4270         4234         4083
##          car       female        woman          man    character
##         3961         3380         2953         2666         2495
##       police         male          sex       father        scene
##         2463         2308         2124         2112         1993
##       nudity       mother        blood         face          son
##         1827         1804         1778         1743         1739
##       murder      stabbed          gun         head          new
##         1699         1692         1689         1653         1636
##       camera         word        child     daughter     american
##         1549         1453         1415         1320         1319
##         film        chase        based         girl         city
##         1317         1316         1312         1311         1292
##         wife          one         hand         york         fire
##         1284         1264         1252         1223         1210
##         year      brother       school        chest        fight
##         1210         1205         1173         1166         1137
##         bare         hero      credits      falling      husband
##         1136         1127         1125         1122         1114
##         dead        black         body     violence    exploding
##         1087         1075         1072         1046         1034
##          dog         love       family        knife          boy
##         1032         1024          986          963          960
##          hit         kiss        phone     someone         name
##          959          945          943          920          878
##    flashback      officer       animal      opening          war
##          869          849          828          825          824
##   friendship         loss       friend        party       sister
##          821          820          816          803          798
##      chested      smoking         back      punched    explosion
##          787          780          770          764          753
##       window       pistol         self     panties      teenage
##          752          743          743          718          714
##   helicopter        train       motion       secret         drug
##          708          699          694          687          683
##          leg       ending     surprise         cell          gay
##          682          681          679          676          676
##        agent      machine       spoken          two      suicide
##          675          666          665          661          657
```

```r
keep.keyword <- c("relationship","death","shot","car","female","woman","man","male","police","sex","nudi
keyword.df <- as.data.frame(keyword.matrix)[keep.keyword]
```

```
head(keyword.df)
```

```
##   relationship death shot car female woman man male police sex nudity
## 1            0     0    0   1      0     0   0    0      0   0      1
## 2            0     0    1   0      1     1   2    0      0   1      0
## 3            8     2    2   0      4     1   2    2      0   0      0
## 4            3     1    0   2      2     0   3    2      0   0      0
## 5            1     2    0   0      3     4   5    4      0   1      2
## 6            2     9    3   8      9     7   4    3      0  11      8
##   blood gun murder hero father mother son american daughter child wife
## 1     0   1      0    1      0      0   0        0        0     0    0
## 2     0   1      0    1      0      0   0        0        0     0    0
## 3     0   0      0    2      1      1   1        0        1     0    0
## 4     0   0      0    1      0      0   0        0        0     3    0
## 5     0   0      0    1      1      6   2        0        0     2    0
## 6     0   1      1    1      0      2   0        2        1     1    1
##   fight school husband violence love family dog war animal friendship
## 1     0      0       0        3    0      0   0   0      0          0
## 2     0      0       0        0    0      0   0   1      0          1
## 3     1      0       0        0    4      0   2   0      8          0
## 4     0      0       0        1    0      0   0   0      2          0
## 5     1      2       0        0    0      1   0   4      0          3
## 6     0      0       0        2    9      0   1   0      0          1
##   teenage
## 1       0
## 2       0
## 3       0
## 4       1
## 5       0
## 6       2
```

4.7 Rating

I don't know why some ratings in corpus cannot be transalted into dtm. Just ignore it for now

```
rating <- as.character(movies.before$rating)
head(rating,10)
```

```
##  [1] "NOT RATED" "PG"        "G"         "PG-13"     "PG-13"
##  [6] "PG-13"     "PG"        "PG-13"     "PG"        "R"
```

```
rating.source <- VectorSource(rating)
rating.corpus <- VCorpus(rating.source)

table(movies.before$rating)
```

```
##
##            APPROVED         G         M       N/A     NC-17 NOT RATED
##         0         2        47         0       142         3        88
##        NR        PG     PG-13         R     TV-14      TV-G     TV-MA
##         1       345       888      1071         0         0         0
##     TV-PG      TV-Y   Unrated   UNRATED         X
##         1         0         0        25         1
```

```
for (i in 1:20){
  print(rating.corpus[[i]][1])
```

```
}
```

```
## $content
## [1] "NOT RATED"
##
## $content
## [1] "PG"
##
## $content
## [1] "G"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG"
##
## $content
## [1] "R"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "R"
##
## $content
## [1] "G"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "R"
##
## $content
```

```
## [1] "PG-13"
##
## $content
## [1] "PG-13"
##
## $content
## [1] "PG-13"
```

```
rating.dtm <- DocumentTermMatrix(rating.corpus)
rating.matrix <- as.matrix(rating.dtm)
dim(rating.matrix)
```

```
## [1] 2614    8
```

```
head(rating.matrix)
```

```
##     Terms
## Docs approved n/a nc-17 not pg-13 rated tv-pg unrated
##    1        0   0     0   1     0     1     0       0
##    2        0   0     0   0     0     0     0       0
##    3        0   0     0   0     0     0     0       0
##    4        0   0     0   0     1     0     0       0
##    5        0   0     0   0     1     0     0       0
##    6        0   0     0   0     1     0     0       0
```

5. Combine these dataframes together

```
dim(movies.clean)
```

```
## [1] 2614    8
```

```
dim(actor.df)
```

```
## [1] 2614   42
```

```
dim(award.df)
```

```
## [1] 2614    4
```

```
dim(director.df)
```

```
## [1] 2614   24
```

```
dim(genre.df)
```

```
## [1] 2614   11
```

```
dim(keyword.df)
```

```
## [1] 2614   33
```

```
dim(production.df)
```

```
## [1] 2614   10
```

```
movies.clean.df <- cbind(movies.clean,actor.df,award.df,director.df,genre.df,keyword.df,production.df)
dim(movies.clean.df)
```

```
## [1] 2614  132
```

```
write.csv(movies.clean.df,"movies_clean.csv")
```

6. split into train and test set for model validation.

Find NAs and in movies.clean.df in imdb_rating, matacritic, imdb_votes, and runtime.

```
summary(movies.clean.df)
```

```
##   imdb_rating      metacritic      imdb_votes          Budget
##  Min.   :1.600   Min.   :  1.0   Min.   :     17   Min.   :4.400e+01
##  1st Qu.:5.800   1st Qu.: 41.0   1st Qu.:   9142   1st Qu.:7.000e+06
##  Median :6.500   Median : 53.0   Median :  38340   Median :2.000e+07
##  Mean   :6.371   Mean   : 53.6   Mean   :  92852   Mean   :6.418e+07
##  3rd Qu.:7.100   3rd Qu.: 66.0   3rd Qu.: 108550   3rd Qu.:5.000e+07
##  Max.   :9.900   Max.   :100.0   Max.   :1827477   Max.   :3.000e+10
##  NA's   :17      NA's   :333     NA's   :17
##  Box.Office.Gross     release.year  release.month       runtime
##  Min.   :      335   Min.   :1950   Min.   : 1.000   Min.   : 41.0
##  1st Qu.:  1243848   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 94.0
##  Median : 18912638   Median :2011   Median : 7.000   Median :104.0
##  Mean   : 46815764   Mean   :2009   Mean   : 6.567   Mean   :107.8
##  3rd Qu.: 58382184   3rd Qu.:2014   3rd Qu.:10.000   3rd Qu.:118.0
##  Max.   :936662225   Max.   :2017   Max.   :12.000   Max.   :226.0
##                                                      NA's   :8
##  zachgalifianakis     willarnett         vindiesel
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.003826   Mean   :0.003826   Mean   :0.004208
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##     tomhanks          tomcruise         stevecarell
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.008416   Mean   :0.004208   Mean   :0.007269
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  scarlettjohansson   ryanreynolds      robertdowneyjr
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004973   Mean   :0.004973   Mean   :0.006886
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##   ralphfiennes      natalieportman       mattdamon
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004973   Mean   :0.004591   Mean   :0.006503
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  matthewmcconaughey  markwahlberg       kristenwiig
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
```

```
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.006503   Mean   :0.006886   Mean   :0.004208
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##    liamneeson       kristenstewart      kirstendunst
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.007651   Mean   :0.003826   Mean   :0.003826
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##    kevinhart          johnnydepp       josephgordonlevitt
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004208   Mean   :0.006121   Mean   :0.004591
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##   johngoodman        jasonbateman      jenniferlawrence
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004591   Mean   :0.006886   Mean   :0.004591
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##    jimcarrey         jamesfranco        hughjackman
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004973   Mean   :0.005356   Mean   :0.003826
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##   harrisonford       gerardbutler       adamsandler
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.003826   Mean   :0.004591   Mean   :0.008034
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##     amyadams        angelinajolie        benstiller
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.006121   Mean   :0.003826   Mean   :0.004973
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  bradleycooper       camerondiaz       cateblanchett
```

```
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.005356   Mean   :0.004973   Mean   :0.004591
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##    chrisevans       chrishemsworth       chrispine
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004973   Mean   :0.004208   Mean   :0.004208
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  dwaynejohnson       eddiemurphy         emmastone          nomination
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:1.0000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :1.0000
##  Mean   :0.006886   Mean   :0.004208   Mean   :0.004208   Mean   :0.9311
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :2.0000
##
##      oscar          golden globe          award          stevenspielberg
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.0000   Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.1461   Mean   :0.04132   Mean   :0.009946   Mean   :0.006886
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##
##   jamescameron     christophernolan     woodyallen
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.00153   Mean   :0.001913   Mean   :0.002295
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##
##    ronhoward        robertzemeckis     stevensoderbergh
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004973   Mean   :0.003443   Mean   :0.003826
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##   dennisdugan          spikelee          oliverstone        peterberg
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.000000   Median :0.000000   Median :0.00000   Median :0.00000
##  Mean   :0.003443   Mean   :0.003443   Mean   :0.00306   Mean   :0.00306
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
```

```
##
##   ridleyscott          anglee            robcohen
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.00306   Mean   :0.002678   Mean   :0.002678
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##
##    timburton           jameswan          michaelbay
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.002678   Mean   :0.002295   Mean   :0.002678
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  clinteastwood        tylerperry        peterjackson
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.002678   Mean   :0.002678   Mean   :0.002295
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##  ethancoen,joelcoen   davidlynch         joewright
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.001913   Mean   :0.001913   Mean   :0.001913
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##
##   sofiacoppola          action         adventure          comedy
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.001913   Mean   :0.2422   Mean   :0.1859   Mean   :0.3692
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##      drama          romance          fantasy            horror
##  Min.   :0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1.000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.505   Mean   :0.1725   Mean   :0.07651   Mean   :0.09373
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##
##     thriller         mystery           scifi             crime
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.000
##  Mean   :0.1618   Mean   :0.0811   Mean   :0.07001   Mean   :0.158
```

```
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000
##
##   relationship        death             shot            car
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 1.000   Median : 0.000   Median : 0.00   Median : 0.000
##  Mean   : 2.147   Mean   : 1.634   Mean   : 1.62   Mean   : 1.515
##  3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 2.00   3rd Qu.: 2.000
##  Max.   :18.000   Max.   :17.000   Max.   :19.00   Max.   :27.000
##
##      female           woman            man             male
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 0.000   Median : 0.00   Median : 0.00   Median : 0.0000
##  Mean   : 1.293   Mean   : 1.13   Mean   : 1.02   Mean   : 0.8829
##  3rd Qu.: 2.000   3rd Qu.: 1.00   3rd Qu.: 1.00   3rd Qu.: 1.0000
##  Max.   :20.000   Max.   :25.00   Max.   :13.00   Max.   :14.0000
##
##      police             sex              nudity            blood
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.9422   Mean   : 0.8125   Mean   : 0.6989   Mean   : 0.6802
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :19.0000   Max.   :24.0000   Max.   :11.0000   Max.   :14.0000
##
##       gun              murder           hero            father
##  Min.   : 0.0000   Min.   : 0.00   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.00   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.00   Median :0.0000   Median : 0.000
##  Mean   : 0.6461   Mean   : 0.65   Mean   :0.4311   Mean   : 0.808
##  3rd Qu.: 1.0000   3rd Qu.: 1.00   3rd Qu.:0.0000   3rd Qu.: 1.000
##  Max.   :13.0000   Max.   :10.00   Max.   :9.0000   Max.   :17.000
##
##      mother            son             american          daughter
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median :0.000
##  Mean   :0.6901   Mean   : 0.6653   Mean   : 0.5046   Mean   :0.505
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.:1.000
##  Max.   :8.0000   Max.   :12.0000   Max.   :14.0000   Max.   :8.000
##
##      child             wife            fight           school
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.: 0.0000
##  Median : 0.0000   Median :0.0000   Median :0.000   Median : 0.0000
##  Mean   : 0.5413   Mean   :0.4912   Mean   :0.435   Mean   : 0.4487
##  3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.: 0.0000
##  Max.   :13.0000   Max.   :6.0000   Max.   :7.000   Max.   :19.0000
##
##     husband           violence          love            family
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
```

```
##  Median : 0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   : 0.4262   Mean   :0.4002   Mean   :0.3917   Mean   :0.3772
##  3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :16.0000   Max.   :8.0000   Max.   :9.0000   Max.   :9.0000
##
##       dog               war             animal          friendship
##  Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 0.0000   Median :0.0000   Median : 0.0000   Median :0.0000
##  Mean   : 0.3948   Mean   :0.3152   Mean   : 0.3168   Mean   :0.3141
##  3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :20.0000   Max.   :8.0000   Max.   :16.0000   Max.   :5.0000
##
##     teenage       universalpictures warnerbros.pictures 20thcenturyfox
##  Min.   :0.0000   Min.   :0.0000    Min.   :0.00000     Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.00000     1st Qu.:0.00000
##  Median :0.0000   Median :0.0000    Median :0.00000     Median :0.00000
##  Mean   :0.2731   Mean   :0.1006    Mean   :0.05356     Mean   :0.05126
##  3rd Qu.:0.0000   3rd Qu.:0.0000    3rd Qu.:0.00000     3rd Qu.:0.00000
##  Max.   :8.0000   Max.   :1.0000    Max.   :1.00000     Max.   :1.00000
##
##   sonypictures    paramountpictures waltdisneypictures focusfeatures
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000     Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000     1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.0000     Median :0.00000
##  Mean   :0.0417   Mean   :0.03673   Mean   :0.0264     Mean   :0.02372
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000     3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000     Max.   :1.00000
##
##  lionsgatefilms    sonypicturesclassics columbiapictures
##  Min.   :0.00000   Min.   :0.00000      Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000      1st Qu.:0.00000
##  Median :0.00000   Median :0.00000      Median :0.00000
##  Mean   :0.01951   Mean   :0.01683      Mean   :0.01262
##  3rd Qu.:0.00000   3rd Qu.:0.00000      3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000      Max.   :1.00000
##
```

```
# Fill in 0 where there's NA
movies.clean.df[which(is.na(movies.clean.df$imdb_rating)),1] <- 0
movies.clean.df[which(is.na(movies.clean.df$metacritic)),2] <- 0
movies.clean.df[which(is.na(movies.clean.df$imdb_votes)),3] <- 0
movies.clean.df[which(is.na(movies.clean.df$runtime)),8] <- 0
```

```
set.seed(1)
train.index <- sample(nrow(movies.clean.df),round(nrow(movies.clean.df)*0.8))
train <- movies.clean.df[train.index,]
test <- movies.clean.df[-train.index,]
```

7. Fit linear regression model

```
train.fit <- lm(Box.Office.Gross~.,data=train)
summary(train.fit)
```

```
##
## Call:
```

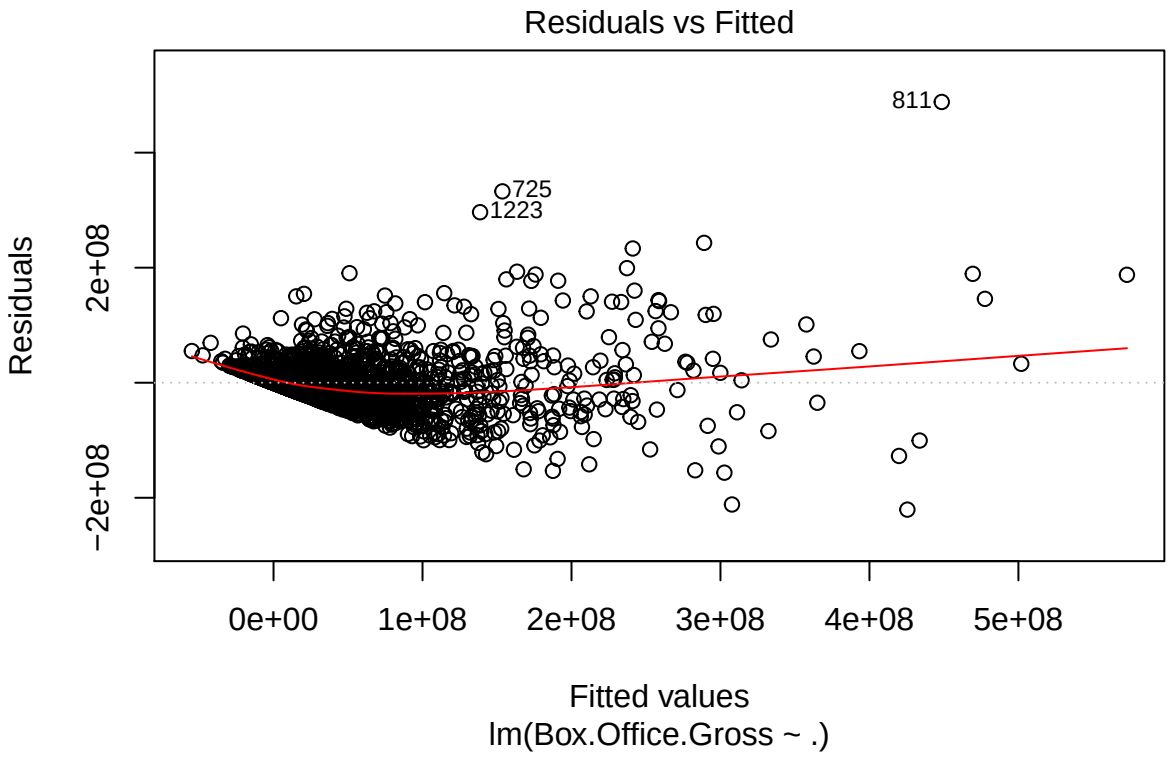
```
## lm(formula = Box.Office.Gross ~ ., data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -220612837  -17673830   -1505238   13532768  488115491
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.080e+08  2.869e+08  -0.376 0.706597
## imdb_rating          -3.490e+06  1.160e+06  -3.010 0.002646 **
## metacritic            8.929e+02  5.555e+04   0.016 0.987177
## imdb_votes            2.562e+02  1.072e+01  23.904  < 2e-16 ***
## Budget                9.449e-04  1.323e-03   0.714 0.475314
## release.year          6.606e+04  1.424e+05   0.464 0.642839
## release.month         1.787e+05  2.887e+05   0.619 0.535950
## runtime               4.916e+04  6.168e+04   0.797 0.425578
## zachgalifianakis     -1.721e+06  1.566e+07  -0.110 0.912504
## willarnett            2.205e+07  1.620e+07   1.361 0.173562
## vindiesel             7.013e+07  1.529e+07   4.587 4.79e-06 ***
## tomhanks             -2.444e+06  1.081e+07  -0.226 0.821181
## tomcruise             1.589e+07  1.611e+07   0.986 0.324026
## stevecarell           2.267e+07  1.333e+07   1.701 0.089097 .
## scarlettjohansson     1.997e+07  1.511e+07   1.321 0.186529
## ryanreynolds          1.244e+07  1.310e+07   0.950 0.342335
## robertdowneyjr        5.825e+07  1.197e+07   4.868 1.22e-06 ***
## ralphfiennes          2.502e+07  1.466e+07   1.706 0.088110 .
## natalieportman        6.139e+06  1.481e+07   0.414 0.678591
## mattdamon             1.133e+07  1.388e+07   0.816 0.414512
## matthewmcconaughey    9.031e+06  1.098e+07   0.822 0.410957
## markwahlberg          1.811e+07  1.312e+07   1.380 0.167730
## kristenwiig           4.752e+07  1.479e+07   3.213 0.001335 **
## liamneeson            6.872e+06  1.116e+07   0.616 0.537954
## kristenstewart        9.936e+07  1.863e+07   5.333 1.08e-07 ***
## kirstendunst          3.777e+07  1.734e+07   2.178 0.029533 *
## kevinhart             6.765e+07  1.593e+07   4.248 2.26e-05 ***
## johnnydepp           -1.032e+07  1.318e+07  -0.783 0.433649
## josephgordonlevitt   -1.245e+07  1.422e+07  -0.875 0.381595
## johngoodman           6.709e+06  1.425e+07   0.471 0.637798
## jasonbateman          1.799e+06  1.191e+07   0.151 0.879930
## jenniferlawrence      4.946e+07  1.635e+07   3.025 0.002515 **
## jimcarrey             1.569e+07  1.442e+07   1.088 0.276539
## jamesfranco           1.029e+07  1.393e+07   0.739 0.460133
## hughjackman           6.856e+06  1.654e+07   0.414 0.678565
## harrisonford          7.840e+07  1.520e+07   5.158 2.75e-07 ***
## gerardbutler          6.030e+06  1.415e+07   0.426 0.670120
## adamsandler           3.325e+07  1.365e+07   2.436 0.014927 *
## amyadams              2.532e+06  1.273e+07   0.199 0.842433
## angelinajolie         1.851e+07  1.544e+07   1.198 0.230876
## benstiller            5.422e+07  1.372e+07   3.953 8.00e-05 ***
## bradleycooper         4.221e+07  1.586e+07   2.662 0.007841 **
## camerondiaz           3.546e+07  1.440e+07   2.463 0.013882 *
## cateblanchett         6.182e+06  1.423e+07   0.434 0.664000
## chrisevans            2.021e+07  1.372e+07   1.474 0.140755
## chrishemsworth       -2.200e+06  1.425e+07  -0.154 0.877307
```

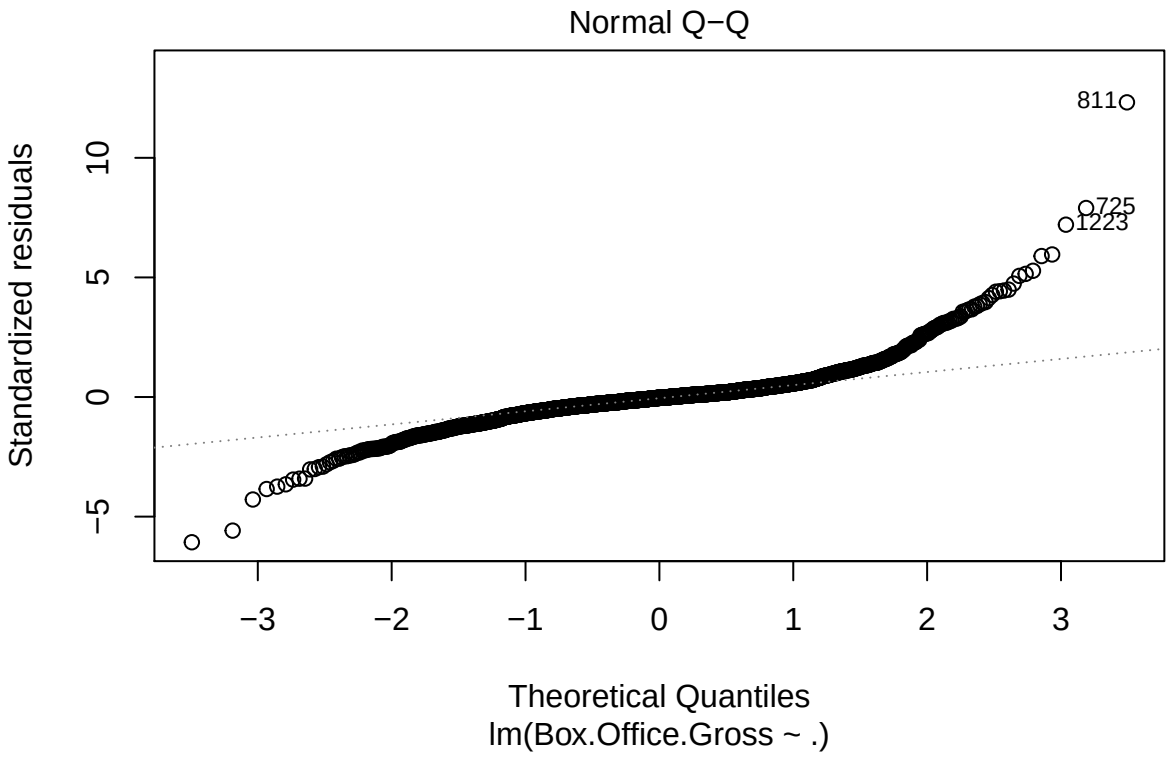
```
## chrispine             1.386e+07  1.593e+07   0.870 0.384618
## dwaynejohnson         1.769e+07  1.186e+07   1.491 0.136103
## eddiemurphy           5.303e+07  1.437e+07   3.690 0.000230 ***
## emmastone             4.521e+06  1.611e+07   0.281 0.779070
## nomination            2.189e+06  2.350e+06   0.932 0.351708
## oscar                 1.869e+07  4.017e+06   4.654 3.47e-06 ***
## `golden globe`        1.149e+07  5.809e+06   1.978 0.048099 *
## award                 6.113e+07  1.045e+07   5.852 5.67e-09 ***
## stevenspielberg       1.747e+07  1.226e+07   1.425 0.154314
## jamescameron          2.063e+08  2.332e+07   8.847  < 2e-16 ***
## christophernolan     -3.778e+07  2.318e+07  -1.630 0.103270
## woodyallen           -1.573e+07  2.038e+07  -0.772 0.440324
## ronhoward             1.935e+07  1.322e+07   1.464 0.143282
## robertzemeckis       -5.862e+06  1.739e+07  -0.337 0.736072
## stevensoderbergh      9.902e+06  1.587e+07   0.624 0.532693
## dennisdugan           1.183e+07  1.888e+07   0.626 0.531135
## spikelee              2.762e+06  1.837e+07   0.150 0.880532
## oliverstone           2.129e+07  1.836e+07   1.159 0.246397
## peterberg             3.450e+07  1.934e+07   1.784 0.074596 .
## ridleyscott          -1.917e+07  2.331e+07  -0.822 0.411002
## anglee               -5.888e+06  1.842e+07  -0.320 0.749237
## robcohen             -4.109e+06  2.023e+07  -0.203 0.839105
## timburton            -1.594e+06  1.797e+07  -0.089 0.929343
## jameswan              6.150e+07  2.277e+07   2.700 0.006984 **
## michaelbay            9.905e+07  1.827e+07   5.421 6.67e-08 ***
## clinteastwood         2.639e+07  1.748e+07   1.509 0.131406
## tylerperry            4.713e+07  1.819e+07   2.591 0.009637 **
## peterjackson          6.081e+07  2.301e+07   2.643 0.008284 **
## `ethancoen,joelcoen` -3.077e+07  2.277e+07  -1.351 0.176738
## davidlynch           -3.387e+07  2.574e+07  -1.316 0.188437
## joewright            -1.975e+07  2.333e+07  -0.846 0.397467
## sofiacoppola         -2.385e+07  2.633e+07  -0.906 0.365252
## action               -4.660e+05  3.213e+06  -0.145 0.884699
## adventure             1.792e+07  3.229e+06   5.551 3.23e-08 ***
## comedy                2.608e+05  2.648e+06   0.098 0.921582
## drama                -1.434e+07  2.576e+06  -5.566 2.96e-08 ***
## romance              -4.597e+05  2.927e+06  -0.157 0.875196
## fantasy              -2.922e+06  3.992e+06  -0.732 0.464320
## horror                2.332e+06  4.235e+06   0.551 0.581986
## thriller             -3.798e+06  3.158e+06  -1.203 0.229218
## mystery               1.528e+06  3.975e+06   0.384 0.700696
## scifi                -6.958e+06  4.379e+06  -1.589 0.112176
## crime                -5.336e+06  3.218e+06  -1.658 0.097414 .
## relationship         -7.153e+05  7.355e+05  -0.973 0.330880
## death                 2.303e+06  7.175e+05   3.209 0.001354 **
## shot                  3.986e+05  5.552e+05   0.718 0.472830
## car                   8.727e+05  4.612e+05   1.892 0.058643 .
## female                4.443e+06  6.787e+05   6.547 7.50e-11 ***
## woman                -6.918e+05  6.930e+05  -0.998 0.318278
## man                  -3.735e+05  9.257e+05  -0.403 0.686682
## male                  2.547e+05  9.156e+05   0.278 0.780933
## police               -3.376e+05  5.715e+05  -0.591 0.554819
## sex                  -7.124e+05  6.431e+05  -1.108 0.268161
## nudity               -4.174e+06  9.997e+05  -4.175 3.11e-05 ***
```

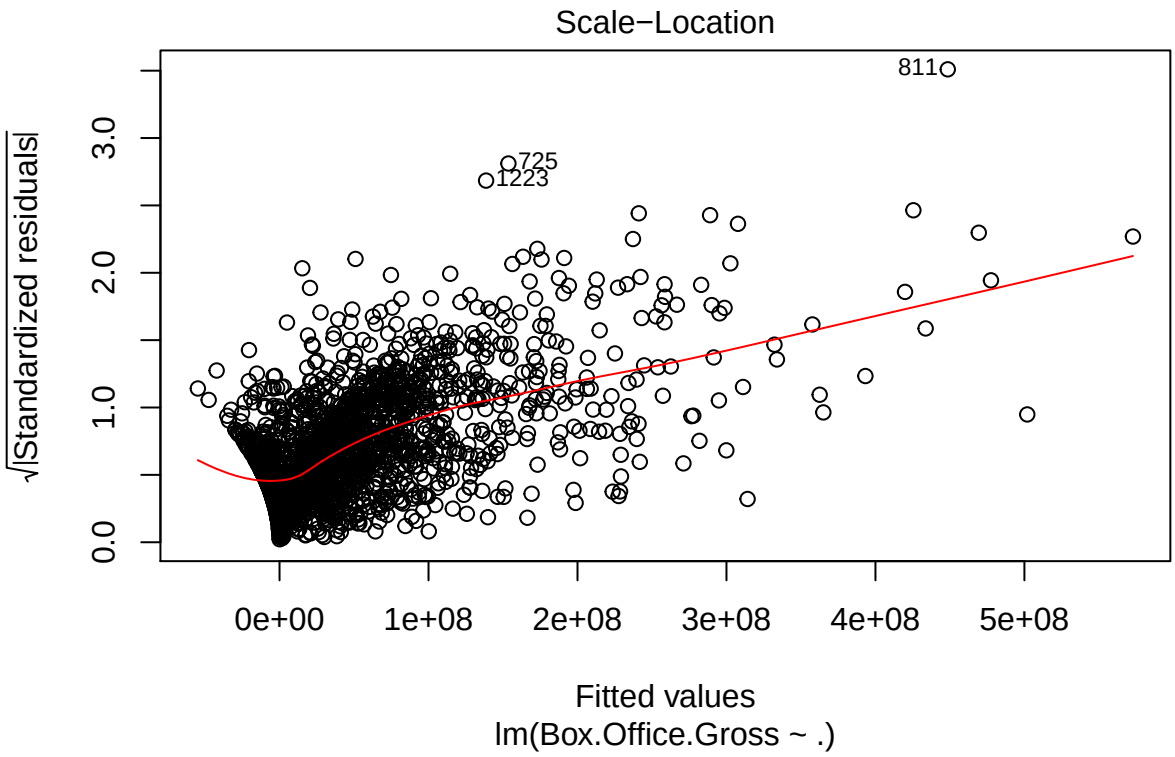
```
## blood                -7.530e+06  1.118e+06  -6.735 2.15e-11 ***
## gun                  -1.641e+06  1.072e+06  -1.532 0.125782
## murder               -1.818e+06  1.304e+06  -1.394 0.163371
## hero                  3.135e+06  1.330e+06   2.357 0.018526 *
## father                2.323e+06  1.238e+06   1.876 0.060740 .
## mother                5.548e+05  1.323e+06   0.419 0.675041
## son                   2.145e+05  1.446e+06   0.148 0.882105
## american              3.895e+05  1.117e+06   0.349 0.727411
## daughter             -8.596e+05  1.558e+06  -0.552 0.581128
## child                -7.974e+05  9.158e+05  -0.871 0.383984
## wife                 -5.033e+05  1.806e+06  -0.279 0.780473
## fight                -2.954e+06  1.424e+06  -2.075 0.038110 *
## school               -1.240e+05  8.392e+05  -0.148 0.882525
## husband              -2.061e+06  1.902e+06  -1.083 0.278770
## violence              3.304e+06  1.799e+06   1.836 0.066445 .
## love                  3.932e+04  1.369e+06   0.029 0.977088
## family                2.984e+06  1.467e+06   2.035 0.042030 *
## dog                   3.741e+05  9.346e+05   0.400 0.689018
## war                  -1.913e+05  1.217e+06  -0.157 0.875102
## animal                5.484e+06  1.149e+06   4.774 1.94e-06 ***
## friendship           -1.345e+06  1.774e+06  -0.758 0.448659
## teenage              -1.841e+06  1.526e+06  -1.207 0.227566
## universalpictures     1.877e+07  3.566e+06   5.263 1.57e-07 ***
## warnerbros.pictures   1.805e+07  4.554e+06   3.964 7.62e-05 ***
## `20thcenturyfox`      9.544e+06  4.802e+06   1.988 0.046978 *
## sonypictures          1.744e+07  5.251e+06   3.322 0.000911 ***
## paramountpictures     8.515e+06  5.381e+06   1.582 0.113746
## waltdisneypictures    7.525e+07  6.695e+06  11.240  < 2e-16 ***
## focusfeatures        -7.988e+06  6.794e+06  -1.176 0.239822
## lionsgatefilms       -4.365e+06  7.176e+06  -0.608 0.543066
## sonypicturesclassics -1.828e+07  7.649e+06  -2.390 0.016930 *
## columbiapictures      2.593e+07  9.351e+06   2.773 0.005611 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43610000 on 1959 degrees of freedom
## Multiple R-squared:  0.703,  Adjusted R-squared:  0.6832
## F-statistic:  35.4 on 131 and 1959 DF,  p-value: < 2.2e-16
```

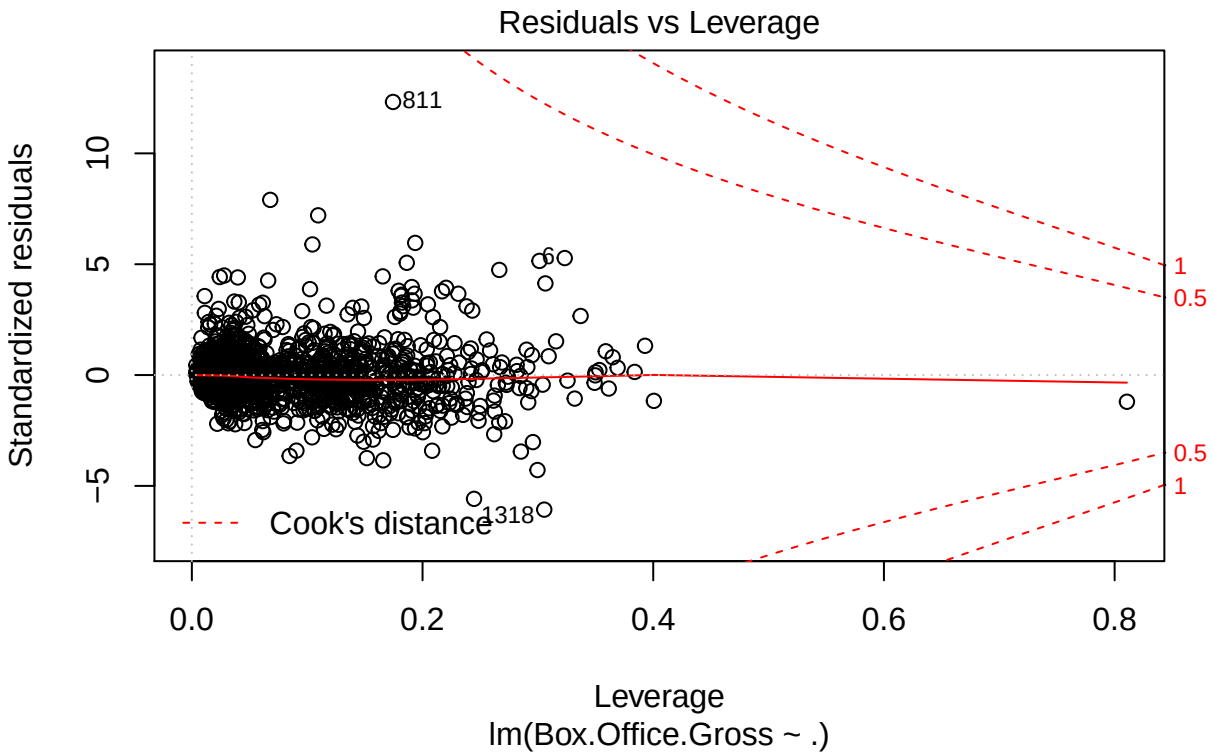
In this model, p-value is significant and we have good explanatory power (68.67% of variability in box office is expalined by these predictors.) given the complexity of the dataset.

```
plot(train.fit)
```


Residuals vs Fitted
811
725
1223
Residuals
2e+08
−2e+08
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
Fitted values
lm(Box.Office.Gross ~ .)


Normal Q−Q
Standardized residuals
10
5
0
−5
811
725
1223
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Box.Office.Gross ~ .)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
811
725
1223
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
Fitted values
lm(Box.Office.Gross ~ .)



We can see that observation 11, 811, 725 tend to be outliers in three plots, so we elimate them from the train set and fit the model again. There's a light improvement but so little that could be ignored.

Predict using test set and calculate rmse.

```
pred <- predict(train.fit,test)
test$pred.lm <- pred
```

There're negative predicted value here, which means our model is not performing very well.

```
pred.result <- test %>%
  select(Box.Office.Gross,pred.lm)
pred.result$ratio.lm <- round(pred.result$pred.lm/pred.result$Box.Office.Gross,2)
head(pred.result)
```

```
##    Box.Office.Gross   pred.lm ratio.lm
## 2         238632124 145721926     0.61
## 11        652270625 144824639     0.22
## 12         13100042  17054195     1.30
## 13         99967670  38934259     0.39
## 14         43818839 117959989     2.69
## 19         77222099 151059693     1.96
```

```
RMSE.lm <- sqrt(mean((test$pred.lm-test$Box.Office.Gross)^2))
RMSE.lm
```

```
## [1] 51503153
```

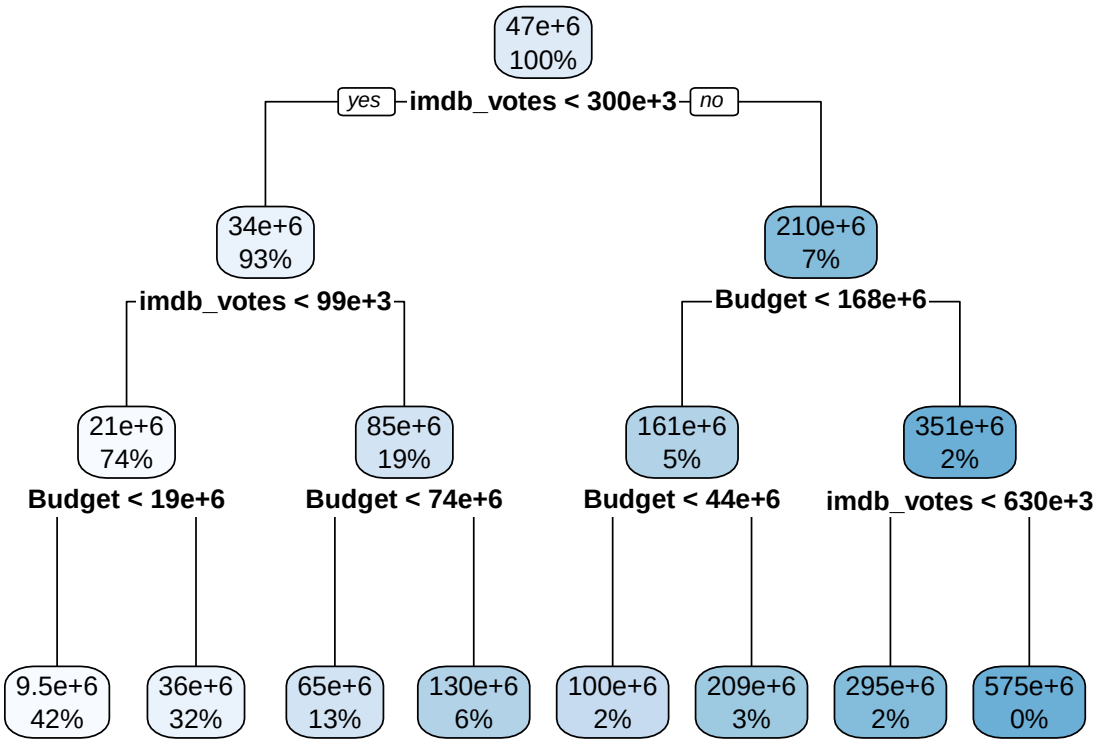
8. Fit regression tree model

```r
library(rpart)
library(rpart.plot)
```

```r
tree.fit <- rpart(formula=Box.Office.Gross~.,data=train,method="anova")
print(tree.fit)
```

```
## n= 2091
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 2091 1.254480e+19  47114260
##    2) imdb_votes< 300139 1936 4.614925e+18  34045430
##      4) imdb_votes< 99345.5 1537 1.442874e+18  20757900
##        8) Budget< 1.875e+07 877 2.536709e+17   9518408 *
##        9) Budget>=1.875e+07 660 9.312013e+17  35692810 *
##      5) imdb_votes>=99345.5 399 1.855327e+18  85230730
##       10) Budget< 7.35e+07 273 6.453383e+17  64527910 *
##       11) Budget>=7.35e+07 126 8.394580e+17 130086800 *
##    3) imdb_votes>=300139 155 3.469194e+18 210348200
##      6) Budget< 1.675e+08 115 1.268800e+18 161453700
##       12) Budget< 4.45e+07 50 2.963738e+17  99531830 *
##       13) Budget>=4.45e+07 65 6.332364e+17 209086000 *
##      7) Budget>=1.675e+08 40 1.135055e+18 350919600
##       14) imdb_votes< 629872.5 32 3.091251e+17 294950200 *
##       15) imdb_votes>=629872.5 8 3.247175e+17 574797200 *
```

```r
rpart.plot(tree.fit)
```




```
pred.tree <- predict(tree.fit,test)

pred.result$pred.tree <- pred.tree
pred.result$ratio.tree <- round(pred.result$pred.tree/pred.result$Box.Office.Gross,2)
head(pred.result)
```

```
##    Box.Office.Gross   pred.lm ratio.lm pred.tree ratio.tree
## 2         238632124 145721926     0.61  64527909       0.27
## 11        652270625 144824639     0.22 209085978       0.32
## 12         13100042  17054195     1.30  35692810       2.72
## 13         99967670  38934259     0.39  35692810       0.36
## 14         43818839 117959989     2.69  64527909       1.47
## 19         77222099 151059693     1.96 130086829       1.68
```

```
RMSE.tree <- sqrt(mean((pred.tree-test$Box.Office.Gross)^2))
RMSE.tree
```

```
## [1] 51786372
```

9. Fit random forest model

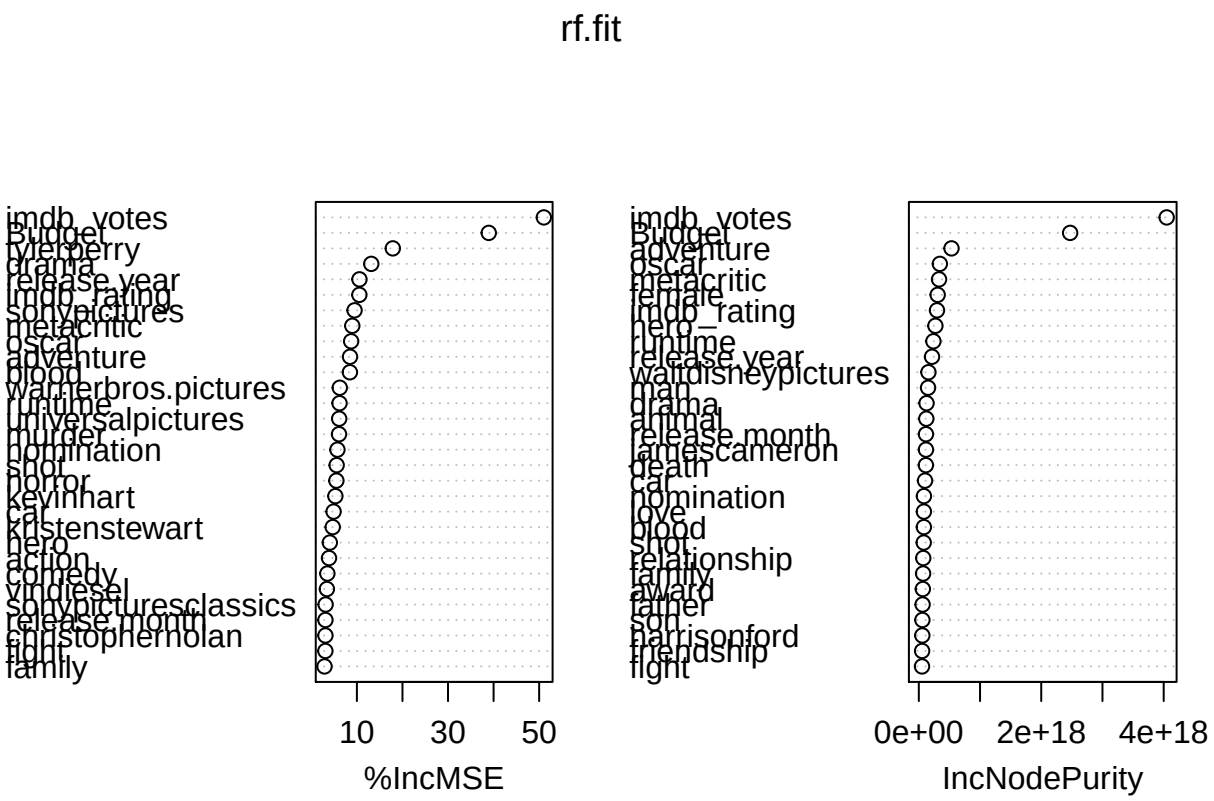
```
library(randomForest)
```

```
names(train) <- make.names(names(train))
rf.fit <- randomForest(formula=Box.Office.Gross~.,data=train,importance=TRUE)
rf.fit
```

```
##
## Call:
```

```
##  randomForest(formula = Box.Office.Gross ~ ., data = train, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 43
##
##           Mean of squared residuals: 1.772134e+15
##                     % Var explained: 70.46
```

```
varImpPlot(rf.fit)
```



70.47% of variance explained, better than linear regression with lower rmse.

```
names(test) <- make.names(names(test))
pred.rf <- predict(rf.fit,test)

pred.result$pred.rf <- pred.rf
pred.result$ratio.rf <- round(pred.rf/pred.result$Box.Office.Gross,2)
head(pred.result)
```

```
##    Box.Office.Gross   pred.lm ratio.lm pred.tree ratio.tree   pred.rf
## 2         238632124 145721926     0.61  64527909       0.27 109346248
## 11        652270625 144824639     0.22 209085978       0.32 198392702
## 12         13100042  17054195     1.30  35692810       2.72  50842791
## 13         99967670  38934259     0.39  35692810       0.36  38998038
## 14         43818839 117959989     2.69  64527909       1.47  72896294
## 19         77222099 151059693     1.96 130086829       1.68 103265281
##    ratio.rf
## 2      0.46
```

```
## 11     0.30
## 12     3.88
## 13     0.39
## 14     1.66
## 19     1.34
```

```
RMSE.rf <- sqrt(mean((pred.rf-test$Box.Office.Gross)^2))
RMSE.rf
```

```
## [1] 43049543
```

10. Fit xgboost

```
library(xgboost)
```

```
xgb.fit <- xgboost(data=as.matrix(train[,-5]),
                   nfold=5,
                   label=as.matrix(train$Box.Office.Gross),
                   nrounds=2200,
                   verbose=FALSE,
                   objective="reg:linear",
                   eval_metric="rmse",
                   nthread=8,
                   eta=0.01,
                   gamma=0.0468,
                   max_depth=15,
                   min_child_weight=1.7817,
                   subsample=0.5213,
                   colsample_bytree=0.4603)
```

```
xgb.fit
```

```
## ##### xgb.Booster
## raw: 25.3 Mb
## call:
##   xgb.train(params = params, data = dtrain, nrounds = nrounds,
##     watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
##     early_stopping_rounds = early_stopping_rounds, maximize = maximize,
##     save_period = save_period, save_name = save_name, xgb_model = xgb_model,
##     callbacks = callbacks, nfold = 5, objective = "reg:linear",
##     eval_metric = "rmse", nthread = 8, eta = 0.01, gamma = 0.0468,
##     max_depth = 15, min_child_weight = 1.7817, subsample = 0.5213,
##     colsample_bytree = 0.4603)
## params (as set within xgb.train):
##   nfold = "5", objective = "reg:linear", eval_metric = "rmse", nthread = "8", eta = "0.01", gamma =
## xgb.attributes:
##   niter
## callbacks:
##   cb.save.model(save_period = save_period, save_name = save_name)
## niter: 2200
```

```
library(Metrics)
```

```
pred.xgb <- predict(xgb.fit,newdata=as.matrix(test[,-5]))

pred.result$pred.xgb <- pred.xgb
pred.result$ratio.xgb <- round(pred.xgb/pred.result$Box.Office.Gross,2)
```

```
head(pred.result)
```

```
##    Box.Office.Gross   pred.lm ratio.lm pred.tree ratio.tree   pred.rf
## 2         238632124 145721926     0.61  64527909       0.27 109346248
## 11        652270625 144824639     0.22 209085978       0.32 198392702
## 12         13100042  17054195     1.30  35692810       2.72  50842791
## 13         99967670  38934259     0.39  35692810       0.36  38998038
## 14         43818839 117959989     2.69  64527909       1.47  72896294
## 19         77222099 151059693     1.96 130086829       1.68 103265281
##    ratio.rf  pred.xgb ratio.xgb
## 2      0.46 105359560      0.44
## 11     0.30 177018816      0.27
## 12     3.88  38901036      2.97
## 13     0.39  32179930      0.32
## 14     1.66  60834128      1.39
## 19     1.34 130545744      1.69
```

```
RMSE.xgb <- rmse(test$Box.Office.Gross,pred.xgb)
RMSE.xgb
```

```
## [1] 41942594
```

xgb has the lowest rmse among the 4 models.