final\_analysis

library(ggplot2)  
library(ggpubr)

## Loading required package: magrittr

library(plyr)

##   
## Attaching package: 'plyr'

## The following object is masked from 'package:ggpubr':  
##   
## mutate

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:plyr':  
##   
## here

## The following object is masked from 'package:base':  
##   
## date

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':  
##   
## intersect, setdiff, union

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

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

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

library(scales)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

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

LOAD DATA

movies <- get(load("movies.RData"))

head(movies)

## title title\_type genre runtime mpaa\_rating  
## 1 Filly Brown Feature Film Drama 80 R  
## 2 The Dish Feature Film Drama 101 PG-13  
## 3 Waiting for Guffman Feature Film Comedy 84 R  
## 4 The Age of Innocence Feature Film Drama 139 PG  
## 5 Malevolence Feature Film Horror 90 R  
## 6 Old Partner Documentary Documentary 78 Unrated  
## studio thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day  
## 1 Indomina Media Inc. 2013 4 19  
## 2 Warner Bros. Pictures 2001 3 14  
## 3 Sony Pictures Classics 1996 8 21  
## 4 Columbia Pictures 1993 10 1  
## 5 Anchor Bay Entertainment 2004 9 10  
## 6 Shcalo Media Group 2009 1 15  
## dvd\_rel\_year dvd\_rel\_month dvd\_rel\_day imdb\_rating imdb\_num\_votes  
## 1 2013 7 30 5.5 899  
## 2 2001 8 28 7.3 12285  
## 3 2001 8 21 7.6 22381  
## 4 2001 11 6 7.2 35096  
## 5 2005 4 19 5.1 2386  
## 6 2010 4 20 7.8 333  
## critics\_rating critics\_score audience\_rating audience\_score  
## 1 Rotten 45 Upright 73  
## 2 Certified Fresh 96 Upright 81  
## 3 Certified Fresh 91 Upright 91  
## 4 Certified Fresh 80 Upright 76  
## 5 Rotten 33 Spilled 27  
## 6 Fresh 91 Upright 86  
## best\_pic\_nom best\_pic\_win best\_actor\_win best\_actress\_win best\_dir\_win  
## 1 no no no no no  
## 2 no no no no no  
## 3 no no no no no  
## 4 no no yes no yes  
## 5 no no no no no  
## 6 no no no no no  
## top200\_box director actor1 actor2  
## 1 no Michael D. Olmos Gina Rodriguez Jenni Rivera  
## 2 no Rob Sitch Sam Neill Kevin Harrington  
## 3 no Christopher Guest Christopher Guest Catherine O'Hara  
## 4 no Martin Scorsese Daniel Day-Lewis Michelle Pfeiffer  
## 5 no Stevan Mena Samantha Dark R. Brandon Johnson  
## 6 no Chung-ryoul Lee Choi Won-kyun Lee Sam-soon  
## actor3 actor4 actor5  
## 1 Lou Diamond Phillips Emilio Rivera Joseph Julian Soria  
## 2 Patrick Warburton Tom Long Genevieve Mooy  
## 3 Parker Posey Eugene Levy Bob Balaban  
## 4 Winona Ryder Richard E. Grant Alec McCowen  
## 5 Brandon Johnson Heather Magee Richard Glover  
## 6 Moo <NA> <NA>  
## imdb\_url  
## 1 http://www.imdb.com/title/tt1869425/  
## 2 http://www.imdb.com/title/tt0205873/  
## 3 http://www.imdb.com/title/tt0118111/  
## 4 http://www.imdb.com/title/tt0106226/  
## 5 http://www.imdb.com/title/tt0388230/  
## 6 http://www.imdb.com/title/tt1334549/  
## rt\_url  
## 1 //www.rottentomatoes.com/m/filly\_brown\_2012/  
## 2 //www.rottentomatoes.com/m/dish/  
## 3 //www.rottentomatoes.com/m/waiting\_for\_guffman/  
## 4 //www.rottentomatoes.com/m/age\_of\_innocence/  
## 5 //www.rottentomatoes.com/m/10004684-malevolence/  
## 6 //www.rottentomatoes.com/m/old-partner/

summary(movies)

## title title\_type genre   
## Length:651 Documentary : 55 Drama :305   
## Class :character Feature Film:591 Comedy : 87   
## Mode :character TV Movie : 5 Action & Adventure: 65   
## Mystery & Suspense: 59   
## Documentary : 52   
## Horror : 23   
## (Other) : 60   
## runtime mpaa\_rating studio   
## Min. : 39.0 G : 19 Paramount Pictures : 37   
## 1st Qu.: 92.0 NC-17 : 2 Warner Bros. Pictures : 30   
## Median :103.0 PG :118 Sony Pictures Home Entertainment: 27   
## Mean :105.8 PG-13 :133 Universal Pictures : 23   
## 3rd Qu.:115.8 R :329 Warner Home Video : 19   
## Max. :267.0 Unrated: 50 (Other) :507   
## NA's :1 NA's : 8   
## thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day dvd\_rel\_year   
## Min. :1970 Min. : 1.00 Min. : 1.00 Min. :1991   
## 1st Qu.:1990 1st Qu.: 4.00 1st Qu.: 7.00 1st Qu.:2001   
## Median :2000 Median : 7.00 Median :15.00 Median :2004   
## Mean :1998 Mean : 6.74 Mean :14.42 Mean :2004   
## 3rd Qu.:2007 3rd Qu.:10.00 3rd Qu.:21.00 3rd Qu.:2008   
## Max. :2014 Max. :12.00 Max. :31.00 Max. :2015   
## NA's :8   
## dvd\_rel\_month dvd\_rel\_day imdb\_rating imdb\_num\_votes   
## Min. : 1.000 Min. : 1.00 Min. :1.900 Min. : 180   
## 1st Qu.: 3.000 1st Qu.: 7.00 1st Qu.:5.900 1st Qu.: 4546   
## Median : 6.000 Median :15.00 Median :6.600 Median : 15116   
## Mean : 6.333 Mean :15.01 Mean :6.493 Mean : 57533   
## 3rd Qu.: 9.000 3rd Qu.:23.00 3rd Qu.:7.300 3rd Qu.: 58300   
## Max. :12.000 Max. :31.00 Max. :9.000 Max. :893008   
## NA's :8 NA's :8   
## critics\_rating critics\_score audience\_rating audience\_score   
## Certified Fresh:135 Min. : 1.00 Spilled:275 Min. :11.00   
## Fresh :209 1st Qu.: 33.00 Upright:376 1st Qu.:46.00   
## Rotten :307 Median : 61.00 Median :65.00   
## Mean : 57.69 Mean :62.36   
## 3rd Qu.: 83.00 3rd Qu.:80.00   
## Max. :100.00 Max. :97.00   
##   
## best\_pic\_nom best\_pic\_win best\_actor\_win best\_actress\_win best\_dir\_win  
## no :629 no :644 no :558 no :579 no :608   
## yes: 22 yes: 7 yes: 93 yes: 72 yes: 43   
##   
##   
##   
##   
##   
## top200\_box director actor1 actor2   
## no :636 Length:651 Length:651 Length:651   
## yes: 15 Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## actor3 actor4 actor5   
## Length:651 Length:651 Length:651   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## imdb\_url rt\_url   
## Length:651 Length:651   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

TRANSFORM DATA

#Runtime:  
movies[!complete.cases(movies['runtime']),]

## title title\_type genre runtime mpaa\_rating studio  
## 334 The End of America Documentary Documentary NA Unrated Indipix  
## thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day dvd\_rel\_year dvd\_rel\_month  
## 334 2008 10 1 2009 1  
## dvd\_rel\_day imdb\_rating imdb\_num\_votes critics\_rating critics\_score  
## 334 20 7.5 739 Fresh 80  
## audience\_rating audience\_score best\_pic\_nom best\_pic\_win  
## 334 Upright 72 no no  
## best\_actor\_win best\_actress\_win best\_dir\_win top200\_box director  
## 334 no no no no Anne Sundberg  
## actor1 actor2 actor3 actor4 actor5  
## 334 Naomi Wolf <NA> <NA> <NA> <NA>  
## imdb\_url  
## 334 http://www.imdb.com/title/tt1294790/  
## rt\_url  
## 334 //www.rottentomatoes.com/m/end\_of\_america/

movies$runtime[movies$title == "The End of America"] <- 74  
#https://www.imdb.com/title/tt1294790/  
  
#Director:  
movies$director[movies$title == "Lorenzo's Oil"] <- "George Miller"  
#https://es.wikipedia.org/wiki/Lorenzo%27s\_Oil\_(pel%C3%ADcula)  
movies$director[movies$title == "The Ninth Gate"] <- "Roman Polanski"  
#https://es.wikipedia.org/wiki/The\_Ninth\_Gate  
  
#Studio:  
movies$studio <- as.character(movies$studio)  
movies$studio[movies$title == "Oliver & Company"] <-  
 "Walt Disney Pictures"  
#https://es.wikipedia.org/wiki/Oliver\_y\_su\_pandilla  
movies$studio[movies$title == "Attack of the 50 Foot Woman"] <-  
 "Woolner Brothers Pictures Inc."  
#https://www.imdb.com/title/tt0051380/  
movies$studio[movies$title == "Inbred"] <- "Melanie Light"  
#https://www.imdb.com/title/tt1723124/fullcredits  
movies$studio[movies$title == "Caveman"] <- "United Artists"  
#https://es.wikipedia.org/wiki/El\_cavern%C3%ADcola  
movies$studio[movies$title == "Dirty Sanchez: The Movie"] <-  
 "Vertigo Films"  
#https://www.rottentomatoes.com/m/dirty\_sanchez  
movies$studio[movies$title == "The Man Who Sued God"] <-  
 "Australian Film Finance Corporation (AFFC), New South Wales Film & Television Office, Showtime Australia See more"  
#https://www.imdb.com/title/tt0268437/  
movies$studio[movies$title == "Inserts"] <-  
 "Film and General Productions"  
#https://www.filmaffinity.com/es/film740308.html  
movies$studio <- factor(movies$studio)  
  
#DVD realease date  
movies[!complete.cases(movies['dvd\_rel\_year']), ]

## title title\_type  
## 100 Charlie: The Life and Art of Charles Chaplin Documentary  
## 184 Streets of Gold Feature Film  
## 261 The Squeeze Feature Film  
## 345 Electric Dreams Feature Film  
## 375 Porky's Revenge Feature Film  
## 377 Teen Wolf Too Feature Film  
## 437 The Last Remake of Beau Geste Feature Film  
## 451 Let It Be Documentary  
## genre runtime mpaa\_rating studio  
## 100 Documentary 132 Unrated Warner Bros.  
## 184 Drama 95 R Live Home Video  
## 261 Action & Adventure 101 PG-13 HBO Video  
## 345 Drama 95 PG MGM  
## 375 Art House & International 92 R 20th Century Fox  
## 377 Science Fiction & Fantasy 95 PG Paramount Home Video  
## 437 Action & Adventure 85 PG MCA Universal Home Video  
## 451 Documentary 81 G United Artists  
## thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day dvd\_rel\_year dvd\_rel\_month  
## 100 2004 2 13 NA NA  
## 184 1986 11 14 NA NA  
## 261 1987 7 10 NA NA  
## 345 1984 7 20 NA NA  
## 375 1985 3 22 NA NA  
## 377 1987 11 20 NA NA  
## 437 1977 7 15 NA NA  
## 451 1970 5 20 NA NA  
## dvd\_rel\_day imdb\_rating imdb\_num\_votes critics\_rating critics\_score  
## 100 NA 8.0 1147 Fresh 95  
## 184 NA 6.0 486 Rotten 31  
## 261 NA 4.5 703 Rotten 17  
## 345 NA 6.5 5149 Rotten 47  
## 375 NA 4.6 5863 Rotten 27  
## 377 NA 3.1 8319 Rotten 7  
## 437 NA 6.0 1680 Rotten 33  
## 451 NA 7.9 3887 Fresh 82  
## audience\_rating audience\_score best\_pic\_nom best\_pic\_win  
## 100 Upright 90 no no  
## 184 Spilled 44 no no  
## 261 Spilled 33 no no  
## 345 Upright 72 no no  
## 375 Spilled 33 no no  
## 377 Spilled 17 no no  
## 437 Upright 65 no no  
## 451 Upright 87 no no  
## best\_actor\_win best\_actress\_win best\_dir\_win top200\_box  
## 100 no no no no  
## 184 no no no no  
## 261 no no no no  
## 345 no no no no  
## 375 no no no no  
## 377 no no no no  
## 437 no no no no  
## 451 no no no no  
## director actor1 actor2  
## 100 Richard Schickel Woody Allen Johnny Depp  
## 184 Joe Roth Klaus Maria Brandauer Adrian Pasdar  
## 261 Roger Young Michael Keaton Rae Dawn Chong  
## 345 Steve Barron Lenny Von Dohlen Virginia Madsen  
## 375 James Komack Dan Monahan Wyatt Knight  
## 377 Christopher Leitch Jason Bateman Kim Darby  
## 437 Marty Feldman Marty Feldman Ann-Margret  
## 451 Michael Lindsay-Hogg Beatles The Paul McCartney  
## actor3 actor4 actor5  
## 100 Sydney Pollack Milos Forman Marcel Marceau  
## 184 Richard Pasdar Wesley Snipes Angela Molina  
## 261 Meat Loaf John Davidson Ric Abernathy  
## 345 Maxwell Caulfield Bud Cort Don Fellows  
## 375 Mark Herrier Tony Ganios Kaki Hunter  
## 377 John Astin Estee Chandler Paul Sand  
## 437 Michael York Peter Ustinov James Earl Jones  
## 451 Ringo Starr George Harrison John Lennon  
## imdb\_url  
## 100 http://www.imdb.com/title/tt0379730/  
## 184 http://www.imdb.com/title/tt0092022/  
## 261 http://www.imdb.com/title/tt0094021/  
## 345 http://www.imdb.com/title/tt0087197/  
## 375 http://www.imdb.com/title/tt0089826/  
## 377 http://www.imdb.com/title/tt0094118/  
## 437 http://www.imdb.com/title/tt0076297/  
## 451 http://www.imdb.com/title/tt0065976/  
## rt\_url  
## 100 //www.rottentomatoes.com/m/charlie\_the\_life\_and\_art\_of\_charles\_chaplin/  
## 184 //www.rottentomatoes.com/m/streets\_of\_gold/  
## 261 //www.rottentomatoes.com/m/1019743-squeeze/  
## 345 //www.rottentomatoes.com/m/1006510-electric\_dreams/  
## 375 //www.rottentomatoes.com/m/porkys\_revenge/  
## 377 //www.rottentomatoes.com/m/teen\_wolf\_too/  
## 437 //www.rottentomatoes.com/m/last\_remake\_of\_beau\_geste/  
## 451 //www.rottentomatoes.com/m/let-it-be/

movies$dvd\_rel\_year[movies$title == "Charlie: The Life and Art of Charles Chaplin"] <-  
 2003  
movies$dvd\_rel\_month[movies$title == "Charlie: The Life and Art of Charles Chaplin"] <-  
 11  
movies$dvd\_rel\_day[movies$title == "Charlie: The Life and Art of Charles Chaplin"] <-  
 5  
#https://www.imdb.com/title/tt0379730/releaseinfo  
movies$dvd\_rel\_year[movies$title == "The Squeeze"] <- 2015  
movies$dvd\_rel\_month[movies$title == "The Squeeze"] <- 6  
movies$dvd\_rel\_day[movies$title == "The Squeeze"] <- 9  
#https://medium.com/@releasebandyal/producer-michael-doven-announces-release-of-the-squeeze-on-dvd-75a982d0a047  
movies$dvd\_rel\_year[movies$title == "Electric Dreams"] <- 1984  
#https://en.wikipedia.org/wiki/Electric\_Dreams\_(film)  
movies$dvd\_rel\_year[movies$title == "The Last Remake of Beau Geste"] <-  
 2010  
movies$dvd\_rel\_month[movies$title == "The Last Remake of Beau Geste"] <-  
 1  
movies$dvd\_rel\_day[movies$title == "The Last Remake of Beau Geste"] <-  
 11  
#https://en.wikipedia.org/wiki/The\_Last\_Remake\_of\_Beau\_Geste  
  
#Actors:  
movies[!complete.cases(movies['actor4']), ]

## title title\_type  
## 6 Old Partner Documentary  
## 25 The Yes Men Fix the World Documentary  
## 131 Africa: The Serengeti Documentary  
## 175 Kicking It Documentary  
## 198 Gaza Strip Documentary  
## 223 Attack of the 50 Foot Woman TV Movie  
## 236 Closet Land Feature Film  
## 334 The End of America Documentary  
## 386 Jonestown: The Life and Death of Peoples Temple Documentary  
## 399 The Last Lions Documentary  
## 454 The Disappearance of Alice Creed Feature Film  
## 552 Saint of 9/11 Documentary  
## 574 Sea Monsters: A Prehistoric Adventure Documentary  
## genre runtime mpaa\_rating  
## 6 Documentary 78 Unrated  
## 25 Documentary 87 Unrated  
## 131 Documentary 39 Unrated  
## 175 Documentary 98 Unrated  
## 198 Documentary 74 Unrated  
## 223 Other 90 R  
## 236 Mystery & Suspense 89 R  
## 334 Documentary 74 Unrated  
## 386 Documentary 86 Unrated  
## 399 Documentary 88 PG  
## 454 Mystery & Suspense 96 R  
## 552 Documentary 90 Unrated  
## 574 Documentary 40 G  
## studio thtr\_rel\_year thtr\_rel\_month  
## 6 Shcalo Media Group 2009 1  
## 25 Cinetic Media 2009 1  
## 131 Houston Museum of Natural Scie 1994 4  
## 175 Liberation Entertainment 2008 6  
## 198 Arab Film Distribution 2002 1  
## 223 Woolner Brothers Pictures Inc. 1993 12  
## 236 Media Home Entertainment 1991 3  
## 334 Indipix 2008 10  
## 386 7th art 2006 10  
## 399 National Geographic Entertainment 2011 2  
## 454 Anchor Bay Entertainment 2010 8  
## 552 IFC 2006 9  
## 574 National Geographic 2007 10  
## thtr\_rel\_day dvd\_rel\_year dvd\_rel\_month dvd\_rel\_day imdb\_rating  
## 6 15 2010 4 20 7.8  
## 25 18 2010 4 1 7.6  
## 131 8 1998 11 18 7.3  
## 175 13 2008 9 9 7.4  
## 198 1 2003 5 6 7.3  
## 223 11 2002 4 2 3.8  
## 236 6 1991 9 12 7.2  
## 334 1 2009 1 20 7.5  
## 386 20 2007 4 10 7.9  
## 399 18 2012 1 3 8.4  
## 454 6 2010 11 23 6.8  
## 552 6 2006 9 12 7.8  
## 574 5 2008 6 24 7.0  
## imdb\_num\_votes critics\_rating critics\_score audience\_rating  
## 6 333 Fresh 91 Upright  
## 25 4541 Certified Fresh 79 Upright  
## 131 535 Fresh 100 Upright  
## 175 325 Rotten 53 Upright  
## 198 285 Fresh 78 Upright  
## 223 2289 Rotten 22 Spilled  
## 236 2098 Rotten 44 Upright  
## 334 739 Fresh 80 Upright  
## 386 3649 Certified Fresh 94 Upright  
## 399 3128 Fresh 87 Upright  
## 454 19937 Certified Fresh 82 Upright  
## 552 180 Fresh 84 Upright  
## 574 723 Fresh 100 Upright  
## audience\_score best\_pic\_nom best\_pic\_win best\_actor\_win  
## 6 86 no no no  
## 25 77 no no no  
## 131 74 no no no  
## 175 71 no no no  
## 198 89 no no no  
## 223 21 no no no  
## 236 86 no no no  
## 334 72 no no no  
## 386 88 no no no  
## 399 89 no no yes  
## 454 67 no no no  
## 552 79 no no no  
## 574 68 no no no  
## best\_actress\_win best\_dir\_win top200\_box director  
## 6 no no no Chung-ryoul Lee  
## 25 no no no Andy Bichlbaum  
## 131 no no no George Casey  
## 175 no no no Jeff Werner  
## 198 no no no James Longley  
## 223 no no no Christopher Guest  
## 236 no no no Radha Bharadwaj  
## 334 no no no Anne Sundberg  
## 386 no no no Stanley Nelson  
## 399 no no no Dereck Joubert  
## 454 no no no J Blakeson  
## 552 no no no Glenn Holsten  
## 574 no no no Sean MacLeod Phillips  
## actor1 actor2 actor3 actor4 actor5  
## 6 Choi Won-kyun Lee Sam-soon Moo <NA> <NA>  
## 25 Andy Bichlbaum Reggie Waits Mike Bonanno <NA> <NA>  
## 131 James Earl Jones <NA> <NA> <NA> <NA>  
## 175 Colin Farrell Brandon Francis <NA> <NA> <NA>  
## 198 <NA> <NA> <NA> <NA> <NA>  
## 223 Daryl Hannah Daniel Baldwin Xander Berkeley <NA> <NA>  
## 236 Madeleine Stowe Alan Rickman <NA> <NA> <NA>  
## 334 Naomi Wolf <NA> <NA> <NA> <NA>  
## 386 <NA> <NA> <NA> <NA> <NA>  
## 399 Jeremy Irons <NA> <NA> <NA> <NA>  
## 454 Gemma Arterton Martin Compston Eddie Marsan <NA> <NA>  
## 552 Ian McKellen <NA> <NA> <NA> <NA>  
## 574 Liev Schreiber <NA> <NA> <NA> <NA>  
## imdb\_url  
## 6 http://www.imdb.com/title/tt1334549/  
## 25 http://www.imdb.com/title/tt1352852/  
## 131 http://www.imdb.com/title/tt0109049/  
## 175 http://www.imdb.com/title/tt1157668/  
## 198 http://www.imdb.com/title/tt0329112/  
## 223 http://www.imdb.com/title/tt0106317/  
## 236 http://www.imdb.com/title/tt0101597/  
## 334 http://www.imdb.com/title/tt1294790/  
## 386 http://www.imdb.com/title/tt0762111/  
## 399 http://www.imdb.com/title/tt1692928/  
## 454 http://www.imdb.com/title/tt1379177/  
## 552 http://www.imdb.com/title/tt0795458/  
## 574 http://www.imdb.com/title/tt1027743/  
## rt\_url  
## 6 //www.rottentomatoes.com/m/old-partner/  
## 25 //www.rottentomatoes.com/m/yes\_men\_fix\_the\_world/  
## 131 //www.rottentomatoes.com/m/imax\_africa\_the\_serengeti/  
## 175 //www.rottentomatoes.com/m/kicking\_it/  
## 198 //www.rottentomatoes.com/m/gaza\_strip/  
## 223 //www.rottentomatoes.com/m/1050445-attack\_of\_the\_50\_foot\_woman/  
## 236 //www.rottentomatoes.com/m/closet\_land/  
## 334 //www.rottentomatoes.com/m/end\_of\_america/  
## 386 //www.rottentomatoes.com/m/jonestown\_the\_life\_and\_death\_of\_peoples\_temple/  
## 399 //www.rottentomatoes.com/m/the\_last\_lions/  
## 454 //www.rottentomatoes.com/m/disappearance\_of\_alice\_creed/  
## 552 //www.rottentomatoes.com/m/saint\_of\_911/  
## 574 //www.rottentomatoes.com/m/sea\_monsters\_a\_prehistoric\_adventure/

movies$actor4[movies$title == "Attack of the 50 Foot Woman"] <-  
 "Roy Gordon"  
#https://www.imdb.com/title/tt0051380/fullcredits/?ref\_=tt\_ov\_st\_sm  
  
movies[!complete.cases(movies['actor5']), ]

## title title\_type  
## 6 Old Partner Documentary  
## 25 The Yes Men Fix the World Documentary  
## 131 Africa: The Serengeti Documentary  
## 175 Kicking It Documentary  
## 198 Gaza Strip Documentary  
## 223 Attack of the 50 Foot Woman TV Movie  
## 236 Closet Land Feature Film  
## 334 The End of America Documentary  
## 386 Jonestown: The Life and Death of Peoples Temple Documentary  
## 399 The Last Lions Documentary  
## 454 The Disappearance of Alice Creed Feature Film  
## 507 My Dinner with Andre Feature Film  
## 552 Saint of 9/11 Documentary  
## 574 Sea Monsters: A Prehistoric Adventure Documentary  
## 609 The Illusionist (L'illusionniste) Feature Film  
## genre runtime mpaa\_rating  
## 6 Documentary 78 Unrated  
## 25 Documentary 87 Unrated  
## 131 Documentary 39 Unrated  
## 175 Documentary 98 Unrated  
## 198 Documentary 74 Unrated  
## 223 Other 90 R  
## 236 Mystery & Suspense 89 R  
## 334 Documentary 74 Unrated  
## 386 Documentary 86 Unrated  
## 399 Documentary 88 PG  
## 454 Mystery & Suspense 96 R  
## 507 Drama 110 PG  
## 552 Documentary 90 Unrated  
## 574 Documentary 40 G  
## 609 Drama 80 PG  
## studio thtr\_rel\_year thtr\_rel\_month  
## 6 Shcalo Media Group 2009 1  
## 25 Cinetic Media 2009 1  
## 131 Houston Museum of Natural Scie 1994 4  
## 175 Liberation Entertainment 2008 6  
## 198 Arab Film Distribution 2002 1  
## 223 Woolner Brothers Pictures Inc. 1993 12  
## 236 Media Home Entertainment 1991 3  
## 334 Indipix 2008 10  
## 386 7th art 2006 10  
## 399 National Geographic Entertainment 2011 2  
## 454 Anchor Bay Entertainment 2010 8  
## 507 New Yorker Films 1981 10  
## 552 IFC 2006 9  
## 574 National Geographic 2007 10  
## 609 Sony Pictures Classics 2010 12  
## thtr\_rel\_day dvd\_rel\_year dvd\_rel\_month dvd\_rel\_day imdb\_rating  
## 6 15 2010 4 20 7.8  
## 25 18 2010 4 1 7.6  
## 131 8 1998 11 18 7.3  
## 175 13 2008 9 9 7.4  
## 198 1 2003 5 6 7.3  
## 223 11 2002 4 2 3.8  
## 236 6 1991 9 12 7.2  
## 334 1 2009 1 20 7.5  
## 386 20 2007 4 10 7.9  
## 399 18 2012 1 3 8.4  
## 454 6 2010 11 23 6.8  
## 507 11 2001 2 13 7.8  
## 552 6 2006 9 12 7.8  
## 574 5 2008 6 24 7.0  
## 609 25 2011 5 10 7.5  
## imdb\_num\_votes critics\_rating critics\_score audience\_rating  
## 6 333 Fresh 91 Upright  
## 25 4541 Certified Fresh 79 Upright  
## 131 535 Fresh 100 Upright  
## 175 325 Rotten 53 Upright  
## 198 285 Fresh 78 Upright  
## 223 2289 Rotten 22 Spilled  
## 236 2098 Rotten 44 Upright  
## 334 739 Fresh 80 Upright  
## 386 3649 Certified Fresh 94 Upright  
## 399 3128 Fresh 87 Upright  
## 454 19937 Certified Fresh 82 Upright  
## 507 10522 Fresh 91 Upright  
## 552 180 Fresh 84 Upright  
## 574 723 Fresh 100 Upright  
## 609 27601 Certified Fresh 90 Upright  
## audience\_score best\_pic\_nom best\_pic\_win best\_actor\_win  
## 6 86 no no no  
## 25 77 no no no  
## 131 74 no no no  
## 175 71 no no no  
## 198 89 no no no  
## 223 21 no no no  
## 236 86 no no no  
## 334 72 no no no  
## 386 88 no no no  
## 399 89 no no yes  
## 454 67 no no no  
## 507 86 no no no  
## 552 79 no no no  
## 574 68 no no no  
## 609 79 no no no  
## best\_actress\_win best\_dir\_win top200\_box director  
## 6 no no no Chung-ryoul Lee  
## 25 no no no Andy Bichlbaum  
## 131 no no no George Casey  
## 175 no no no Jeff Werner  
## 198 no no no James Longley  
## 223 no no no Christopher Guest  
## 236 no no no Radha Bharadwaj  
## 334 no no no Anne Sundberg  
## 386 no no no Stanley Nelson  
## 399 no no no Dereck Joubert  
## 454 no no no J Blakeson  
## 507 no no no Louis Malle  
## 552 no no no Glenn Holsten  
## 574 no no no Sean MacLeod Phillips  
## 609 no no no Sylvain Chomet  
## actor1 actor2 actor3 actor4  
## 6 Choi Won-kyun Lee Sam-soon Moo <NA>  
## 25 Andy Bichlbaum Reggie Waits Mike Bonanno <NA>  
## 131 James Earl Jones <NA> <NA> <NA>  
## 175 Colin Farrell Brandon Francis <NA> <NA>  
## 198 <NA> <NA> <NA> <NA>  
## 223 Daryl Hannah Daniel Baldwin Xander Berkeley Roy Gordon  
## 236 Madeleine Stowe Alan Rickman <NA> <NA>  
## 334 Naomi Wolf <NA> <NA> <NA>  
## 386 <NA> <NA> <NA> <NA>  
## 399 Jeremy Irons <NA> <NA> <NA>  
## 454 Gemma Arterton Martin Compston Eddie Marsan <NA>  
## 507 Jean Lenauer Roy Butler Andre Gregory Wallace Shawn  
## 552 Ian McKellen <NA> <NA> <NA>  
## 574 Liev Schreiber <NA> <NA> <NA>  
## 609 Jean-Claude Donda Eilidh Rankin Duncan MacNeil Raymond Mearns  
## actor5 imdb\_url  
## 6 <NA> http://www.imdb.com/title/tt1334549/  
## 25 <NA> http://www.imdb.com/title/tt1352852/  
## 131 <NA> http://www.imdb.com/title/tt0109049/  
## 175 <NA> http://www.imdb.com/title/tt1157668/  
## 198 <NA> http://www.imdb.com/title/tt0329112/  
## 223 <NA> http://www.imdb.com/title/tt0106317/  
## 236 <NA> http://www.imdb.com/title/tt0101597/  
## 334 <NA> http://www.imdb.com/title/tt1294790/  
## 386 <NA> http://www.imdb.com/title/tt0762111/  
## 399 <NA> http://www.imdb.com/title/tt1692928/  
## 454 <NA> http://www.imdb.com/title/tt1379177/  
## 507 <NA> http://www.imdb.com/title/tt0082783/  
## 552 <NA> http://www.imdb.com/title/tt0795458/  
## 574 <NA> http://www.imdb.com/title/tt1027743/  
## 609 <NA> http://www.imdb.com/title/tt0775489/  
## rt\_url  
## 6 //www.rottentomatoes.com/m/old-partner/  
## 25 //www.rottentomatoes.com/m/yes\_men\_fix\_the\_world/  
## 131 //www.rottentomatoes.com/m/imax\_africa\_the\_serengeti/  
## 175 //www.rottentomatoes.com/m/kicking\_it/  
## 198 //www.rottentomatoes.com/m/gaza\_strip/  
## 223 //www.rottentomatoes.com/m/1050445-attack\_of\_the\_50\_foot\_woman/  
## 236 //www.rottentomatoes.com/m/closet\_land/  
## 334 //www.rottentomatoes.com/m/end\_of\_america/  
## 386 //www.rottentomatoes.com/m/jonestown\_the\_life\_and\_death\_of\_peoples\_temple/  
## 399 //www.rottentomatoes.com/m/the\_last\_lions/  
## 454 //www.rottentomatoes.com/m/disappearance\_of\_alice\_creed/  
## 507 //www.rottentomatoes.com/m/my\_dinner\_with\_andre/  
## 552 //www.rottentomatoes.com/m/saint\_of\_911/  
## 574 //www.rottentomatoes.com/m/sea\_monsters\_a\_prehistoric\_adventure/  
## 609 //www.rottentomatoes.com/m/the\_illusionist-2009/

movies$actor5[movies$title == "Attack of the 50 Foot Woman"] <-  
 "George Douglas"  
#https://www.imdb.com/title/tt0051380/fullcredits/?ref\_=tt\_ov\_st\_sm  
movies$actor5[movies$title == "The Illusionist (L'illusionniste)"] <-  
 "Eleanor Tomlinson"  
#https://en.wikipedia.org/wiki/The\_Illusionist\_(2006\_film)  
  
  
#Column manipulation  
movies$thtr\_rel\_date <- as.Date(paste0(movies$thtr\_rel\_year, "-", movies$thtr\_rel\_month, "-", movies$thtr\_rel\_day))  
movies$dvd\_rel\_date <- as.Date(paste0(movies$dvd\_rel\_year, "-", movies$dvd\_rel\_month, "-", movies$dvd\_rel\_day))  
  
movies$thtr\_rel\_decade <- as.numeric(format(movies$thtr\_rel\_date,"%Y")) - (as.numeric(format(movies$thtr\_rel\_date,"%Y")) %% 10)  
movies$thtr\_rel\_decade <- as.factor(movies$thtr\_rel\_decade)  
  
movies$top200\_box <- NULL  
  
movies$dvd\_rel\_day <- NULL  
movies$dvd\_rel\_month <- NULL  
movies$dvd\_rel\_year <- NULL  
  
#movies$thtr\_rel\_day <- NULL  
#movies$thtr\_rel\_month <- NULL  
#movies$thtr\_rel\_year <- NULL  
  
movies$imdb\_url <- NULL  
movies$rt\_url <- NULL

head(movies)

## title title\_type genre runtime mpaa\_rating  
## 1 Filly Brown Feature Film Drama 80 R  
## 2 The Dish Feature Film Drama 101 PG-13  
## 3 Waiting for Guffman Feature Film Comedy 84 R  
## 4 The Age of Innocence Feature Film Drama 139 PG  
## 5 Malevolence Feature Film Horror 90 R  
## 6 Old Partner Documentary Documentary 78 Unrated  
## studio thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day  
## 1 Indomina Media Inc. 2013 4 19  
## 2 Warner Bros. Pictures 2001 3 14  
## 3 Sony Pictures Classics 1996 8 21  
## 4 Columbia Pictures 1993 10 1  
## 5 Anchor Bay Entertainment 2004 9 10  
## 6 Shcalo Media Group 2009 1 15  
## imdb\_rating imdb\_num\_votes critics\_rating critics\_score audience\_rating  
## 1 5.5 899 Rotten 45 Upright  
## 2 7.3 12285 Certified Fresh 96 Upright  
## 3 7.6 22381 Certified Fresh 91 Upright  
## 4 7.2 35096 Certified Fresh 80 Upright  
## 5 5.1 2386 Rotten 33 Spilled  
## 6 7.8 333 Fresh 91 Upright  
## audience\_score best\_pic\_nom best\_pic\_win best\_actor\_win best\_actress\_win  
## 1 73 no no no no  
## 2 81 no no no no  
## 3 91 no no no no  
## 4 76 no no yes no  
## 5 27 no no no no  
## 6 86 no no no no  
## best\_dir\_win director actor1 actor2  
## 1 no Michael D. Olmos Gina Rodriguez Jenni Rivera  
## 2 no Rob Sitch Sam Neill Kevin Harrington  
## 3 no Christopher Guest Christopher Guest Catherine O'Hara  
## 4 yes Martin Scorsese Daniel Day-Lewis Michelle Pfeiffer  
## 5 no Stevan Mena Samantha Dark R. Brandon Johnson  
## 6 no Chung-ryoul Lee Choi Won-kyun Lee Sam-soon  
## actor3 actor4 actor5 thtr\_rel\_date  
## 1 Lou Diamond Phillips Emilio Rivera Joseph Julian Soria 2013-04-19  
## 2 Patrick Warburton Tom Long Genevieve Mooy 2001-03-14  
## 3 Parker Posey Eugene Levy Bob Balaban 1996-08-21  
## 4 Winona Ryder Richard E. Grant Alec McCowen 1993-10-01  
## 5 Brandon Johnson Heather Magee Richard Glover 2004-09-10  
## 6 Moo <NA> <NA> 2009-01-15  
## dvd\_rel\_date thtr\_rel\_decade  
## 1 2013-07-30 2010  
## 2 2001-08-28 2000  
## 3 2001-08-21 1990  
## 4 2001-11-06 1990  
## 5 2005-04-19 2000  
## 6 2010-04-20 2000

summary(movies)

## title title\_type genre   
## Length:651 Documentary : 55 Drama :305   
## Class :character Feature Film:591 Comedy : 87   
## Mode :character TV Movie : 5 Action & Adventure: 65   
## Mystery & Suspense: 59   
## Documentary : 52   
## Horror : 23   
## (Other) : 60   
## runtime mpaa\_rating studio   
## Min. : 39.0 G : 19 Paramount Pictures : 37   
## 1st Qu.: 92.0 NC-17 : 2 Warner Bros. Pictures : 30   
## Median :103.0 PG :118 Sony Pictures Home Entertainment: 27   
## Mean :105.8 PG-13 :133 Universal Pictures : 23   
## 3rd Qu.:115.5 R :329 Warner Home Video : 19   
## Max. :267.0 Unrated: 50 (Other) :514   
## NA's : 1   
## thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day imdb\_rating   
## Min. :1970 Min. : 1.00 Min. : 1.00 Min. :1.900   
## 1st Qu.:1990 1st Qu.: 4.00 1st Qu.: 7.00 1st Qu.:5.900   
## Median :2000 Median : 7.00 Median :15.00 Median :6.600   
## Mean :1998 Mean : 6.74 Mean :14.42 Mean :6.493   
## 3rd Qu.:2007 3rd Qu.:10.00 3rd Qu.:21.00 3rd Qu.:7.300   
## Max. :2014 Max. :12.00 Max. :31.00 Max. :9.000   
##   
## imdb\_num\_votes critics\_rating critics\_score audience\_rating  
## Min. : 180 Certified Fresh:135 Min. : 1.00 Spilled:275   
## 1st Qu.: 4546 Fresh :209 1st Qu.: 33.00 Upright:376   
## Median : 15116 Rotten :307 Median : 61.00   
## Mean : 57533 Mean : 57.69   
## 3rd Qu.: 58300 3rd Qu.: 83.00   
## Max. :893008 Max. :100.00   
##   
## audience\_score best\_pic\_nom best\_pic\_win best\_actor\_win best\_actress\_win  
## Min. :11.00 no :629 no :644 no :558 no :579   
## 1st Qu.:46.00 yes: 22 yes: 7 yes: 93 yes: 72   
## Median :65.00   
## Mean :62.36   
## 3rd Qu.:80.00   
## Max. :97.00   
##   
## best\_dir\_win director actor1 actor2   
## no :608 Length:651 Length:651 Length:651   
## yes: 43 Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## actor3 actor4 actor5   
## Length:651 Length:651 Length:651   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## thtr\_rel\_date dvd\_rel\_date thtr\_rel\_decade  
## Min. :1970-05-20 Min. :1991-03-28 1970: 51   
## 1st Qu.:1990-12-07 1st Qu.:2001-05-15 1980:103   
## Median :2000-09-15 Median :2004-03-02 1990:161   
## Mean :1998-06-15 Mean :2004-12-06 2000:234   
## 3rd Qu.:2007-05-17 3rd Qu.:2008-02-19 2010:102   
## Max. :2014-12-25 Max. :2015-06-09   
## NA's :5

# SDA Homework 1.1

We chose the “Films” dataset and we are going to answer the following questions that we thought were interesting.

### 1. Is there a relationship between the type/genre and score?

### 2. Comparing different scores (critics/audience/imdb)

\* is there any difference on the audience/critics score by genre?

### 3. Comparison best picture/oscar wins and score/oscar wins

\* are oscar-awarded films more liked?

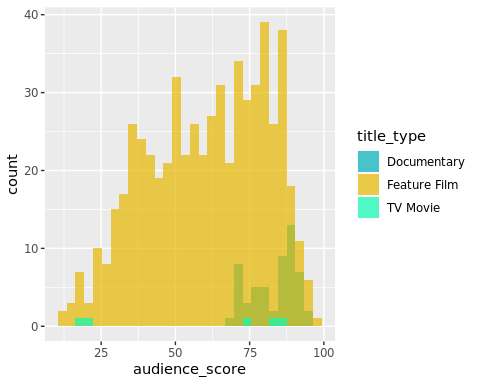
### 4. What are the trends over the years?

\* Preferred month for releases? number of films by month (over the years? does it affect the score?)  
\* Does the genre change over the years?  
\* Do older films tend to have higher score/number of votes? if this is true, it could be because only good old films get remembered

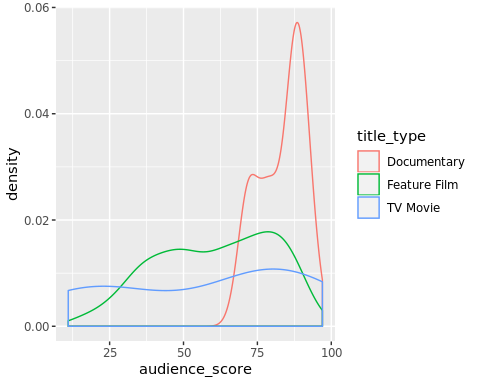
### 5. Do actors and directors have a higher score once they won? (find the first film and see the tendency, look for sudden change)

## 1. Is there a relationship between the type/genre and score?

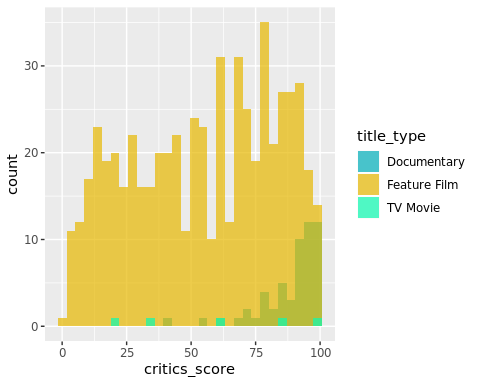
ggplot(data=movies, aes(x=audience\_score)) + geom\_histogram(bins=30, aes(fill = title\_type), position = "identity", alpha = 0.7) + scale\_color\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0")) + scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0"))



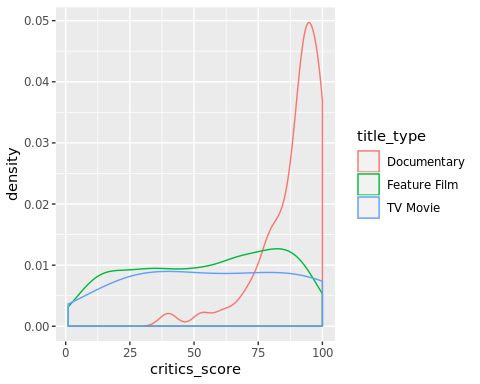
ggplot(data=movies, aes(x=audience\_score)) + geom\_density(aes(color = title\_type), kernel = "gaussian")



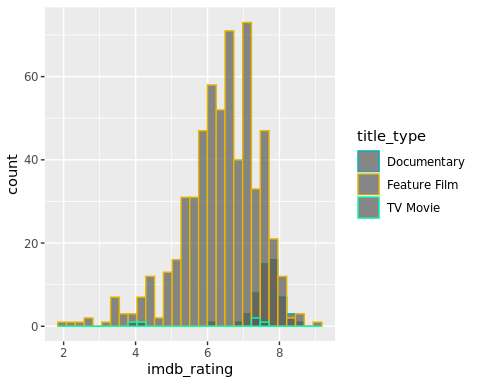
ggplot(data=movies, aes(x=critics\_score)) + geom\_histogram(bins=30, aes(fill = title\_type), position = "identity", alpha = 0.7) + scale\_color\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0")) + scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0"))



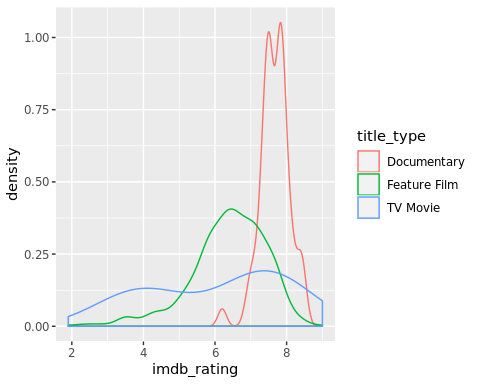
ggplot(data=movies, aes(x=critics\_score)) + geom\_density(aes(color = title\_type), kernel = "gaussian")



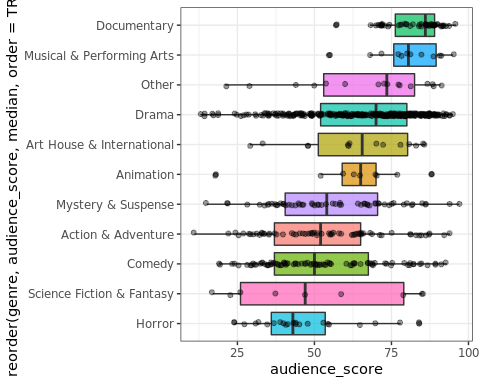
ggplot(data=movies, aes(x=imdb\_rating)) + geom\_histogram(bins=30, aes(color = title\_type), position = "identity", alpha = 0.7) + scale\_color\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0")) + scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#0AFBB0"))



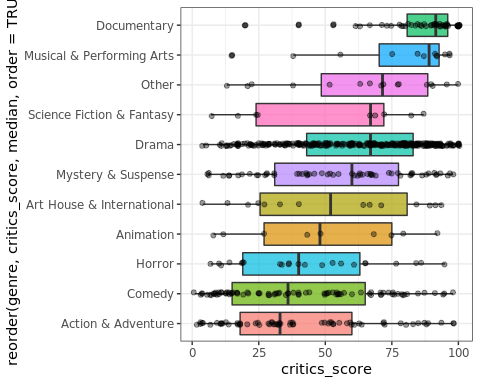
ggplot(data=movies, aes(x=imdb\_rating)) + geom\_density(aes(color = title\_type), kernel = "gaussian")



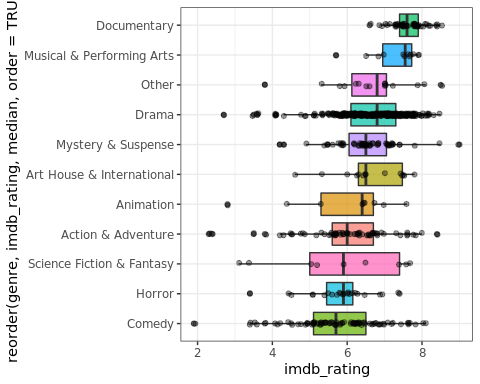
ggplot(data=movies, aes(x = reorder(genre, audience\_score, median, order=TRUE), y = audience\_score, group = genre, fill = genre)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw() + coord\_flip()



ggplot(data=movies, aes(x = reorder(genre, critics\_score, median, order=TRUE), y = critics\_score, group = genre, fill = genre)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw() + coord\_flip()



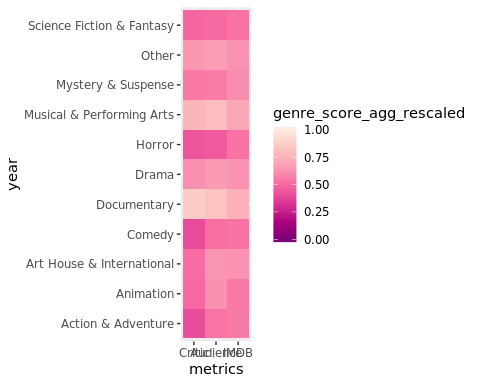
ggplot(data=movies, aes(x = reorder(genre, imdb\_rating, median, order=TRUE), y = imdb\_rating, group = genre, fill = genre)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw() + coord\_flip()



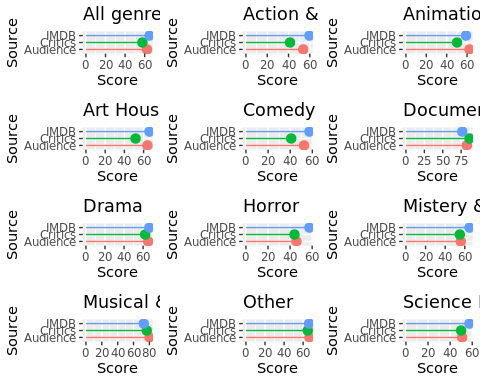
## 2. Comparing different scores (critics/audience/imdb)

### - is there any difference on the audience/critics score by genre?

score = list();  
  
score$genre\_score\_agg = list(  
 "critics" = aggregate(critics\_score ~ genre , movies, mean),  
 "audience" = aggregate(audience\_score ~ genre, movies, mean),  
 "imdb" = aggregate(imdb\_rating ~ genre, movies, mean)  
)  
  
  
score$genre\_metrics\_grid = expand.grid(year =sort(unique(movies$genre)),  
 metrics = c("Critic", "Audience", "IMDB"))  
  
score$genre\_metrics\_grid$genre\_score\_agg =c(score$genre\_score\_agg$critics$critics\_score,  
 score$genre\_score\_agg$audience$audience\_score,   
 score$genre\_score\_agg$imdb$imdb\_rating)  
  
score$genre\_metrics\_grid$genre\_score\_agg\_rescaled =   
 c(  
 rescale(score$genre\_score\_agg$critics$critics\_score, from=(c(1,100))),   
 rescale(score$genre\_score\_agg$audience$audience\_score, from=(c(1,100))),   
 rescale(score$genre\_score\_agg$imdb$imdb\_rating, from=(c(1,10))))  
  
ggplot(score$genre\_metrics\_grid)+  
 geom\_tile(aes(x=metrics, y=year,fill=genre\_score\_agg\_rescaled))+  
 scale\_fill\_distiller( limits=c(0,1), palette = "RdPu")



plot\_number\_movies\_per\_genre <-  
 qplot(movies$genre, fill = movies$genre) + theme(legend.position = "none",  
 axis.text.x = element\_text(angle = 45, hjust = 1)) + ggtitle("Number of movies per genre") +  
 xlab("Genre") + ylab("Number of movies")  
  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(movies$imdb\_rating)\*10,  
 mean(movies$audience\_score),  
 mean(movies$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_all <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("All genres")  
  
#Action and adventure  
action\_adventure <- subset(movies, genre == "Action & Adventure")  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(action\_adventure$imdb\_rating)\*10,  
 mean(action\_adventure$audience\_score),  
 mean(action\_adventure$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_action <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Action & Adventure")  
  
#Animation  
animation <- subset(movies, genre == "Animation")  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(animation$imdb\_rating)\*10,  
 mean(animation$audience\_score),  
 mean(animation$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_animation <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Animation")  
  
#Art House & International  
art\_house <- subset(movies, genre == "Art House & International")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(art\_house$imdb\_rating)\*10,  
 mean(art\_house$audience\_score),  
 mean(art\_house$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_art\_house <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Art House & International")  
  
#Comedy  
comedy <- subset(movies, genre == "Comedy")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(comedy$imdb\_rating)\*10,  
 mean(comedy$audience\_score),  
 mean(comedy$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
plot\_comedy <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Comedy")  
  
#Documentary  
documentary <- subset(movies, genre == "Documentary")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(documentary$imdb\_rating)\*10,  
 mean(documentary$audience\_score),  
 mean(documentary$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_documentary <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Documentary")  
  
#Drama  
drama <- subset(movies, genre == "Drama")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(drama$imdb\_rating)\*10,  
 mean(drama$audience\_score),  
 mean(drama$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_drama <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Drama")  
  
#Horror  
horror <- subset(movies, genre == "Horror")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(horror$imdb\_rating)\*10,  
 mean(horror$audience\_score),  
 mean(horror$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_horror <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Horror")  
  
#Musical & Performing Arts  
musical <- subset(movies, genre == "Musical & Performing Arts")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(musical$imdb\_rating)\*10,  
 mean(musical$audience\_score),  
 mean(musical$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_musical <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Musical & Performing Arts")  
  
#Mystery & Suspense  
mistery <- subset(movies, genre == "Mystery & Suspense")  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(mistery$imdb\_rating)\*10,  
 mean(mistery$audience\_score),  
 mean(mistery$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
  
plot\_mistery <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Mistery & Suspense")  
  
#Other  
other <- subset(movies, genre == "Other")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(other$imdb\_rating)\*10,  
 mean(other$audience\_score),  
 mean(other$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
plot\_other <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Other")  
  
#Science Fiction & Fantasy  
science <- subset(movies, genre == "Science Fiction & Fantasy")  
  
mean\_score <-  
 data.frame(  
 "Score" = c(  
 mean(science$imdb\_rating)\*10,  
 mean(science$audience\_score),  
 mean(science$critics\_score)  
 ),  
 "Source" = c("IMDB", "Audience", "Critics")  
 )  
plot\_science <- ggplot(mean\_score) + geom\_linerange(aes(  
 x = Source,  
 ymin = 0,  
 ymax = Score,  
 colour = Source  
),  
position = position\_dodge(width = 1)) + geom\_point(  
 aes(x = Source, y = Score, colour = Source),  
 position = position\_dodge(width = 1),  
 size = 3  
) + coord\_flip() + theme(legend.position = "none") + ggtitle("Science Fiction & Fantasy")  
  
#Combine the plots...  
plot\_combined\_genre <-  
 ggarrange(  
 plot\_all,  
 plot\_action,  
 plot\_animation,  
 plot\_art\_house,  
 plot\_comedy,  
 plot\_documentary,  
 plot\_drama,  
 plot\_horror,  
 plot\_mistery,  
 plot\_musical,  
 plot\_other,  
 plot\_science,  
 ncol = 3,  
 nrow = 4  
 )  
  
plot\_combined\_genre



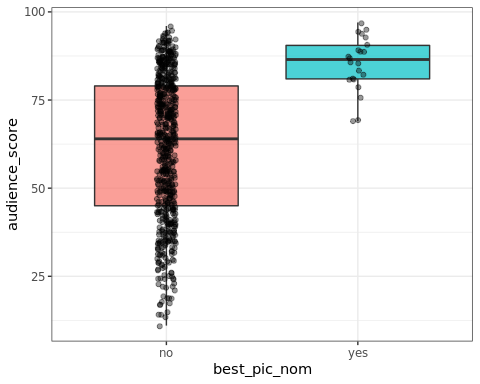
Taking a look at the different lollipop charts, we can see different things: \* Audience gives a higher score than Critics in 10/11 genres (90.9%). The only exception is the Documentary genre. \* IMDB gives a higher score than Critics in 9/11 genres (81.8%). The only exception is the Documentary and Musical & Performance genres. \* IMDB gives a higher score than Audience in 7/11 genres (63.6%). \* Genres like Documentary and Other receive similar ratings from all IMDB, Critics and Audience, whereas in genres like Comedy or Action & Adventure the ratings differ more.

Taking this into account, we can say that the pattern observed in previously is maintained in the individual genres. Critics give the lowest ratings for almost all genres and IMDB and Audience alternate in giving the highest score. There is a difference in the raitings of the different sources according to the genre and this difference follows a pattern, which is more accentuated in some genres than in others.

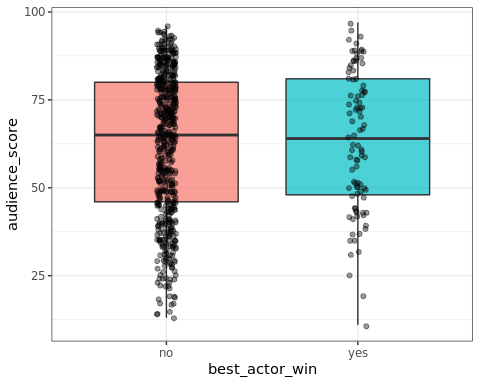
## 3. Comparison best picture/oscar nominated and score/oscar nominated

### - are oscar-awarded films more liked?

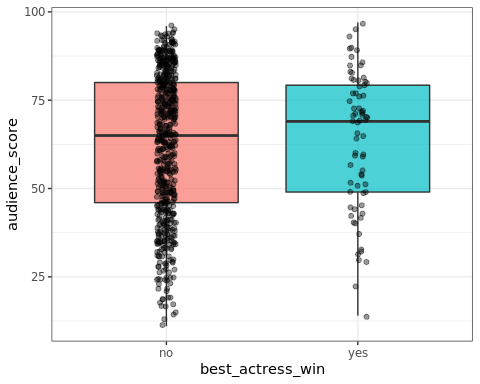
ggplot(data=movies, aes(x = best\_pic\_nom, y = audience\_score, group = best\_pic\_nom, fill = best\_pic\_nom)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



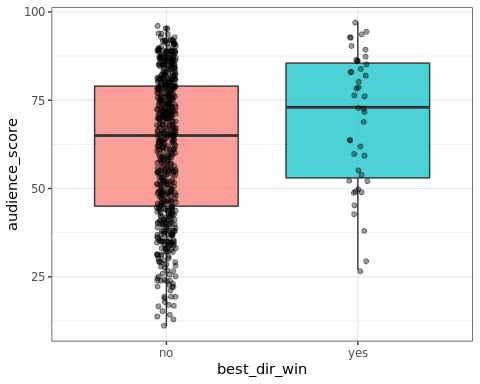
ggplot(data=movies, aes(x = best\_actor\_win, y = audience\_score, group = best\_actor\_win, fill = best\_actor\_win)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



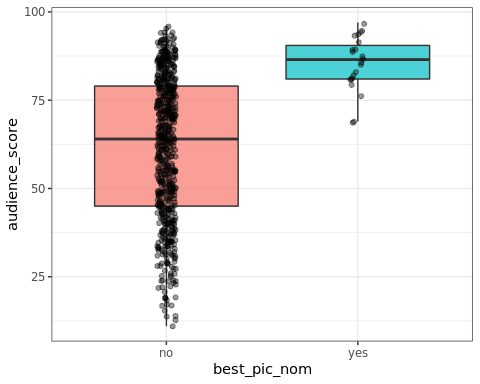
ggplot(data=movies, aes(x = best\_actress\_win, y = audience\_score, group = best\_actress\_win, fill = best\_actress\_win)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



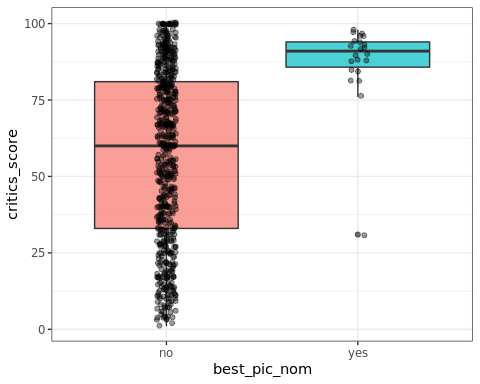
ggplot(data=movies, aes(x = best\_dir\_win, y = audience\_score, group = best\_dir\_win, fill = best\_dir\_win)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



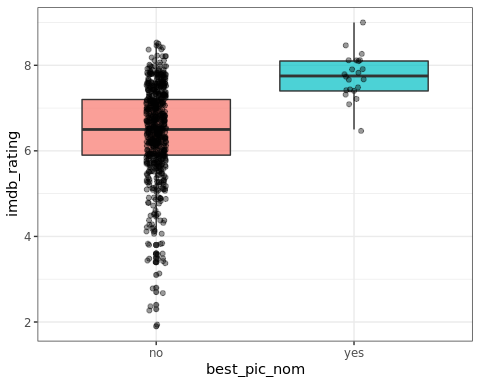
ggplot(data=movies, aes(x = best\_pic\_nom, y = audience\_score, group = best\_pic\_nom, fill = best\_pic\_nom)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



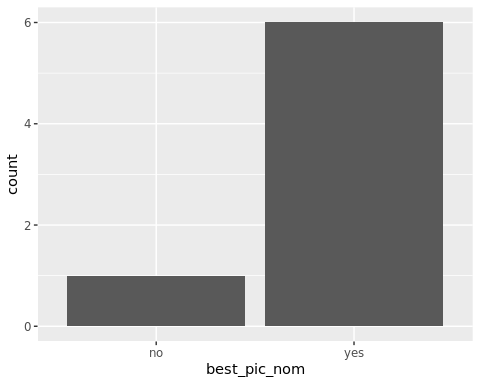
ggplot(data=movies, aes(x = best\_pic\_nom, y = critics\_score, group = best\_pic\_nom, fill = best\_pic\_nom)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



ggplot(data=movies, aes(x = best\_pic\_nom, y = imdb\_rating, group = best\_pic\_nom, fill = best\_pic\_nom)) + geom\_boxplot(alpha = .7) + geom\_jitter(width = .05, alpha = .4) + guides(fill = "none") + theme\_bw()



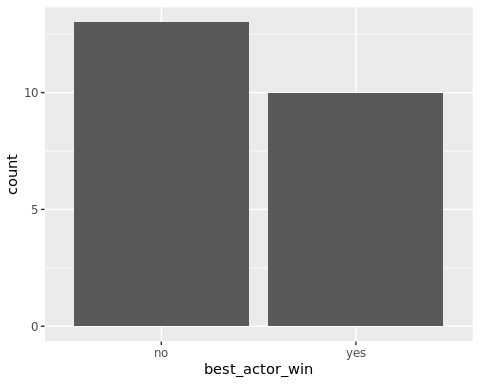
ggplot(aes(x=best\_pic\_nom), data=movies[movies$best\_pic\_win=="yes",]) + geom\_bar()



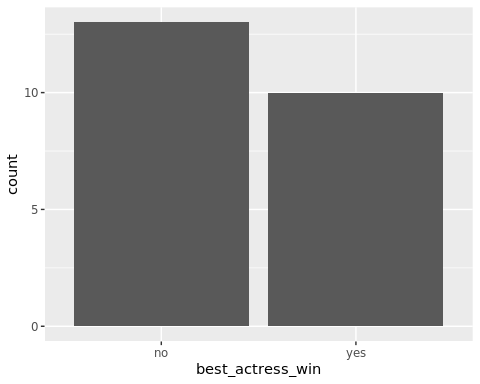
movies[movies$best\_pic\_nom=="no" & movies$best\_pic\_win=="yes",]

## title title\_type genre runtime mpaa\_rating  
## 382 The Hurt Locker Feature Film Drama 131 R  
## studio thtr\_rel\_year thtr\_rel\_month thtr\_rel\_day  
## 382 Summit Entertainment 2009 6 26  
## imdb\_rating imdb\_num\_votes critics\_rating critics\_score  
## 382 7.6 318019 Certified Fresh 98  
## audience\_rating audience\_score best\_pic\_nom best\_pic\_win  
## 382 Upright 84 no yes  
## best\_actor\_win best\_actress\_win best\_dir\_win director  
## 382 no no yes Kathryn Bigelow  
## actor1 actor2 actor3 actor4 actor5  
## 382 Jeremy Renner Anthony Mackie Brian Geraghty Guy Pearce Ralph Fiennes  
## thtr\_rel\_date dvd\_rel\_date thtr\_rel\_decade  
## 382 2009-06-26 2010-01-12 2000

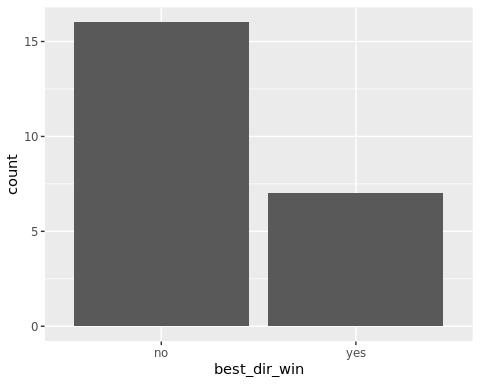
movies[movies$best\_pic\_nom=="no" & movies$best\_pic\_win=="yes",]$best\_pic\_nom <- "yes"  
ggplot(aes(x=best\_actor\_win), data=movies[movies$best\_pic\_nom=="yes",]) + geom\_bar()



ggplot(aes(x=best\_actress\_win), data=movies[movies$best\_pic\_nom=="yes",]) + geom\_bar()



ggplot(aes(x=best\_dir\_win), data=movies[movies$best\_pic\_nom=="yes",]) + geom\_bar()



best\_pic\_movies <- subset(movies, c(best\_pic\_win == "yes"))  
  
 #Add a column showing if they have won another award  
 best\_pic\_movies$other\_win <- NA  
  
 if (best\_pic\_movies$best\_actor\_win == "yes" ||  
 best\_pic\_movies$best\_actress\_win == "yes" ||  
 best\_pic\_movies$best\_dir\_win == "yes") {  
 best\_pic\_movies$other\_win <- "yes"  
 } else{  
 best\_pic\_movies$other\_win <- "no"  
 }  
  
 #Plots...  
 #Best movies that have won another award  
 best\_movie\_another\_award <- plyr::count(best\_pic\_movies$other\_win)  
 best\_movie\_another\_award <- best\_movie\_another\_award %>%  
 arrange((freq)) %>%  
 mutate(lab.ypos = cumsum(freq) - 0.5 \* freq)  
  
 pie\_another\_award <-  
 ggplot(best\_movie\_another\_award, aes(x = "", y = freq, fill = x)) +  
 geom\_bar(width = 1,  
 stat = "identity",  
 color = "white") +  
 coord\_polar("y", start = 0) +  
 theme\_void() +  
 geom\_text(aes(label = signif((freq / 7) \* 100), digits = 2),  
 position = position\_stack(vjust = 0.5),  
 color = "white") +  
 labs(fill = "Another award") +  
 ggtitle("Best movies with another award") +  
 theme(plot.title = element\_text(size = 20, hjust = 0.5)) +  
 scale\_fill\_manual(values = "#00BFC4")

## Warning: Ignoring unknown aesthetics: digits

#Best movies that have won best actor award  
 best\_movie\_best\_actor <- plyr::count(best\_pic\_movies$best\_actor\_win)  
 best\_movie\_best\_actor <- best\_movie\_best\_actor %>%  
 arrange((freq)) %>%  
 mutate(lab.ypos = cumsum(freq) - 0.5 \* freq)  
  
 pie\_best\_actor <-  
 ggplot(best\_movie\_best\_actor, aes(x = "", y = freq, fill = x)) +  
 geom\_bar(width = 1,  
 stat = "identity",  
 color = "white") +  
 coord\_polar("y", start = 0) +  
 theme\_void() +  
 geom\_text(aes(label = signif((freq / 7) \* 100), digits = 2),  
 position = position\_stack(vjust = 0.5),  
 color = "white") +  
 labs(fill = "Best actor") +  
 ggtitle("Best movies with best actor") +  
 theme(plot.title = element\_text(size = 20, hjust = 0.5))

## Warning: Ignoring unknown aesthetics: digits

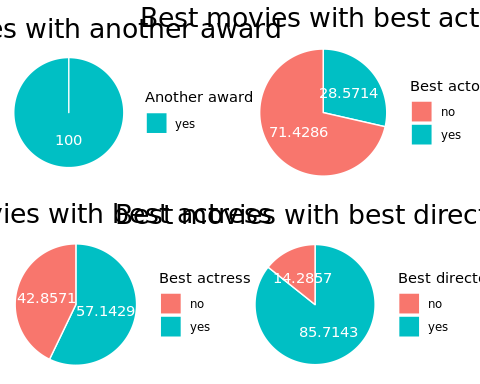
#Best movies that have won best actress award  
 best\_movie\_best\_actress <- plyr::count(best\_pic\_movies$best\_actress\_win)  
 best\_movie\_best\_actress <- best\_movie\_best\_actress %>%  
 arrange((freq)) %>%  
 mutate(lab.ypos = cumsum(freq) - 1 \* freq)  
  
 pie\_best\_actress <-  
 ggplot(best\_movie\_best\_actress, aes(x = "", y = freq, fill = x)) +  
 geom\_bar(width = 1,  
 stat = "identity",  
 color = "white") +  
 coord\_polar("y", start = 0) +  
 theme\_void() +  
 geom\_text(aes(label = signif((freq / 7) \* 100), digits = 2),  
 position = position\_stack(vjust = 0.5),  
 color = "white") +  
 labs(fill = "Best actress") +  
 ggtitle("Best movies with best actress") +  
 theme(plot.title = element\_text(size = 20, hjust = 0.5))

## Warning: Ignoring unknown aesthetics: digits

#Best movies that have won best director award  
 best\_movie\_best\_director <- plyr::count(best\_pic\_movies$best\_dir\_win)  
 best\_movie\_best\_director <- best\_movie\_best\_director %>%  
 arrange((freq)) %>%  
 mutate(lab.ypos = cumsum(freq) - 0.9 \* freq)  
  
 pie\_best\_director <-  
 ggplot(best\_movie\_best\_director, aes(x = "", y = freq, fill = x)) +  
 geom\_bar(width = 1,  
 stat = "identity",  
 color = "white") +  
 coord\_polar("y", start = 0) +  
 theme\_void() +  
 geom\_text(aes(label = signif((freq / 7) \* 100), digits = 2),  
 position = position\_stack(vjust = 0.5),  
 color = "white") +  
 labs(fill = "Best director") +  
 ggtitle("Best movies with best director") +  
 theme(plot.title = element\_text(size = 20, hjust = 0.5))

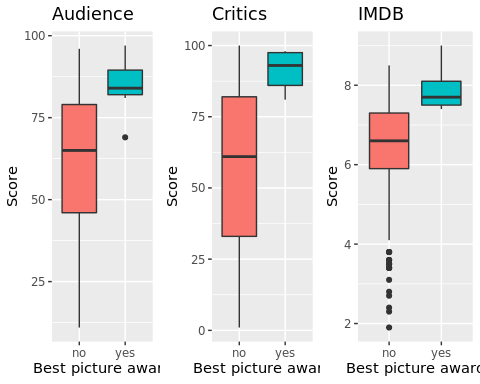
## Warning: Ignoring unknown aesthetics: digits

#Combine plots!  
 plot\_combined\_awarded <-  
 ggarrange(  
 pie\_another\_award,  
 pie\_best\_actor,  
 pie\_best\_actress,  
 pie\_best\_director,  
 ncol = 2,  
 nrow = 2  
 )  
  
plot\_combined\_awarded

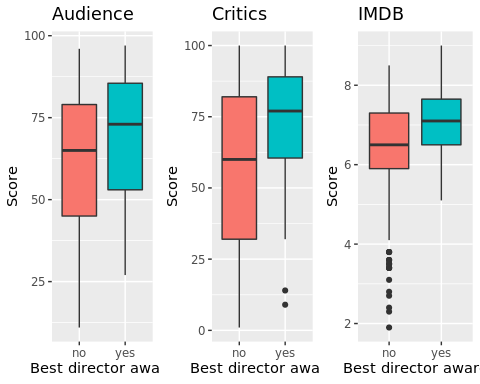


Looking at the pie charts, we can see that the movies in the dataset that have received the award for best film always receive another award. The award for best director is present in 6/7 movies, hinting that the quality of the director is mor influent than that of the actors and actresses in determining if the movie is awarded for best film.

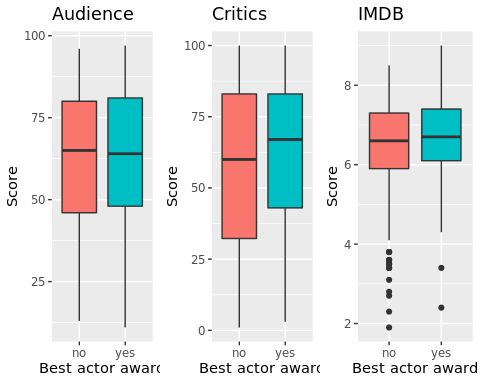
# -are oscar-awarded films more liked?  
#Audience score vs Best\_pic\_win  
plot\_audience\_vs\_best\_pic <-  
 ggplot(movies,  
 aes(x = best\_pic\_win, y = audience\_score, fill = best\_pic\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "Audience", x = "Best picture award", y = "Score")  
#Critics score vs Best\_pic\_win  
plot\_critics\_vs\_best\_pic <-  
 ggplot(movies,  
 aes(x = best\_pic\_win, y = critics\_score, fill = best\_pic\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "Critics", x = "Best picture award", y = "Score")  
#IMDB score vs Best\_pic\_win  
plot\_imdb\_vs\_best\_pic <-  
 ggplot(movies, aes(x = best\_pic\_win, y = imdb\_rating, fill = best\_pic\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "IMDB", x = "Best picture award", y = "Score")  
#Combine the plots...  
plot\_combined\_best\_pic <-  
 ggarrange(  
 plot\_audience\_vs\_best\_pic,  
 plot\_critics\_vs\_best\_pic,  
 plot\_imdb\_vs\_best\_pic,  
 ncol = 3,  
 nrow = 1  
 )  
plot\_combined\_best\_pic



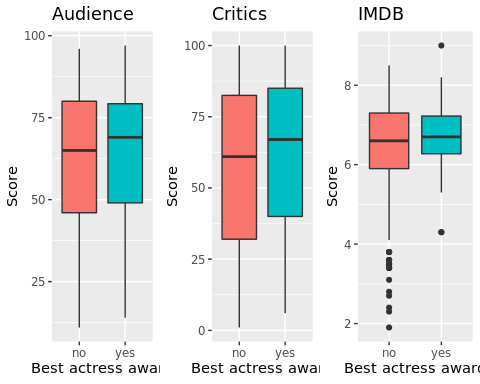
#Audience score vs Best\_dir\_win  
plot\_audience\_vs\_best\_dir <-  
 ggplot(movies,  
 aes(x = best\_dir\_win, y = audience\_score, fill = best\_dir\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "Audience", x = "Best director award", y = "Score")  
#Critics score vs Best\_dir\_win  
plot\_critics\_vs\_best\_dir <-  
 ggplot(movies,  
 aes(x = best\_dir\_win, y = critics\_score, fill = best\_dir\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "Critics", x = "Best director award", y = "Score")  
#IMDB score vs Best\_dir\_win  
plot\_imdb\_vs\_best\_dir <-  
 ggplot(movies, aes(x = best\_dir\_win, y = imdb\_rating, fill = best\_dir\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "IMDB", x = "Best director award", y = "Score")  
#Combine the plots...  
plot\_combined\_best\_dir <-  
 ggarrange(  
 plot\_audience\_vs\_best\_dir,  
 plot\_critics\_vs\_best\_dir,  
 plot\_imdb\_vs\_best\_dir,  
 ncol = 3,  
 nrow = 1  
 )  
plot\_combined\_best\_dir



#Audience score vs Best\_actor\_win  
plot\_audience\_vs\_best\_actor <-  
 ggplot(movies,  
 aes(x = best\_actor\_win, y = audience\_score, fill = best\_actor\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "Audience", x = "Best actor award", y = "Score")  
#Critics score vs Best\_dir\_win  
plot\_critics\_vs\_best\_actor <-  
 ggplot(movies,  
 aes(x = best\_actor\_win, y = critics\_score, fill = best\_actor\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "Critics", x = "Best actor award", y = "Score")  
#IMDB score vs Best\_dir\_win  
plot\_imdb\_vs\_best\_actor <-  
 ggplot(movies,  
 aes(x = best\_actor\_win, y = imdb\_rating, fill = best\_actor\_win)) + geom\_boxplot() + theme(legend.position = "none") + labs(title = "IMDB", x = "Best actor award", y = "Score")  
#Combine the plots...  
plot\_combined\_best\_actor <-  
 ggarrange(  
 plot\_audience\_vs\_best\_actor,  
 plot\_critics\_vs\_best\_actor,  
 plot\_imdb\_vs\_best\_actor,  
 ncol = 3,  
 nrow = 1  
 )  
plot\_combined\_best\_actor



#Audience score vs Best\_actress\_win  
plot\_audience\_vs\_best\_actress <-  
 ggplot(movies,  
 aes(x = best\_actress\_win, y = audience\_score, fill = best\_actress\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "Audience", x = "Best actress award", y = "Score")  
#Critics score vs Best\_dir\_win  
plot\_critics\_vs\_best\_actress <-  
 ggplot(movies,  
 aes(x = best\_actress\_win, y = critics\_score, fill = best\_actress\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "Critics", x = "Best actress award", y = "Score")  
#IMDB score vs Best\_dir\_win  
plot\_imdb\_vs\_best\_actress <-  
 ggplot(movies,  
 aes(x = best\_actress\_win, y = imdb\_rating, fill = best\_actress\_win)) +  
 geom\_boxplot() + theme(legend.position = "none") + labs(title = "IMDB", x = "Best actress award", y = "Score")  
#Combine the plots...  
plot\_combined\_best\_actress <-  
 ggarrange(  
 plot\_audience\_vs\_best\_actress,  
 plot\_critics\_vs\_best\_actress,  
 plot\_imdb\_vs\_best\_actress,  
 ncol = 3,  
 nrow = 1  
 )  
plot\_combined\_best\_actress

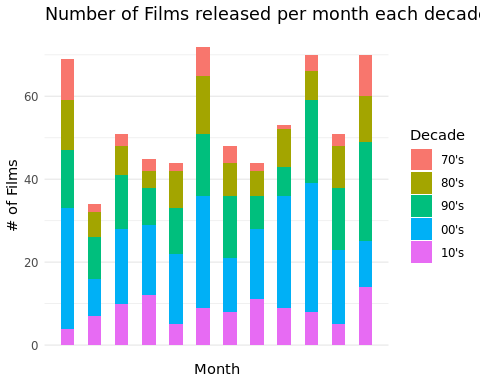


## 4. What are the trends over the years?

- Preferred month for releases? number of films by month (over the years? does it affect the score?)  
- Does the genre change over the years?  
- Do older films tend to have higher score/number of votes? if this is true, it could be because only good old films get remembered

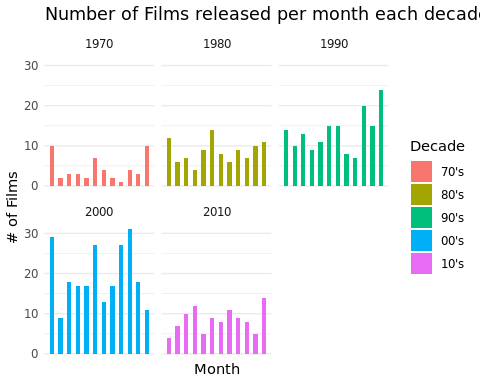
We start to count the movies per months to see if there is a pattern in all this time in the release months.

ggplot(movies, aes(fill = thtr\_rel\_decade)) + geom\_bar(aes(x = thtr\_rel\_month), width = 0.5) + theme\_minimal() + xlab("Month") + ylab("# of Films") + ggtitle("Number of Films released per month each decade ") + scale\_fill\_discrete(name = "Decade", labels=c("70's","80's","90's","00's","10's")) + scale\_x\_discrete(breaks=1:12,labels=c("Jan", "Feb", "Mar", "Apr","May", "Jun", "Jul", "Aug", "Sept", "Oct", "Nov", "Dec"))



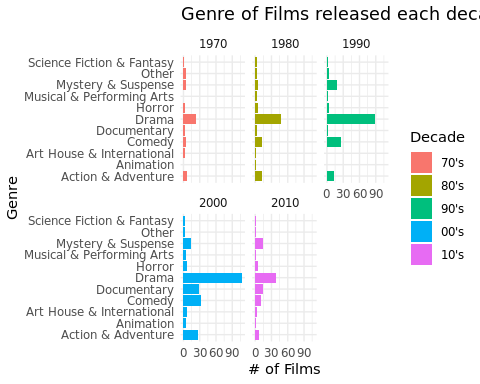
We can see that overall there is a “pattern” of releases in the months of January, June, October and December, so let’s take a better look at this data by decade. We breakdown the previous graphs grouping the data by decade.

ggplot(movies, aes(fill = thtr\_rel\_decade)) + geom\_bar(aes(x = thtr\_rel\_month), width = 0.5) + facet\_wrap(thtr\_rel\_decade ~ .) + theme\_minimal() + xlab("Month") + ylab("# of Films") + ggtitle("Number of Films released per month each decade w/ breakdown") + scale\_fill\_discrete(name = "Decade", labels=c("70's","80's","90's","00's","10's")) + scale\_x\_discrete(breaks=1:12,labels=c("Jan", "Feb", "Mar", "Apr","May", "Jun", "Jul", "Aug", "Sept", "Oct", "Nov", "Dec"))

 > Now we can see clearly that since the 70’s there has been a trend of two moments in the years to release movies in theaters: in summer (June) and in the winter Holidays (December and January). However, in this last decade, we only see a spike in December.

### - Does the genre change over the years (by decade)?

ggplot(movies, aes(fill = thtr\_rel\_decade)) + geom\_bar(aes(x = genre)) + facet\_wrap(thtr\_rel\_decade ~ .) + theme\_minimal() + xlab("Genre") + ylab("# of Films") + ggtitle("Genre of Films released each decade w/ breakdown") + scale\_fill\_discrete(name = "Decade", labels=c("70's","80's","90's","00's","10's")) + coord\_flip()



We can see an increment over the years in the Documentary Genre, as well as in Mystery and Suspense.

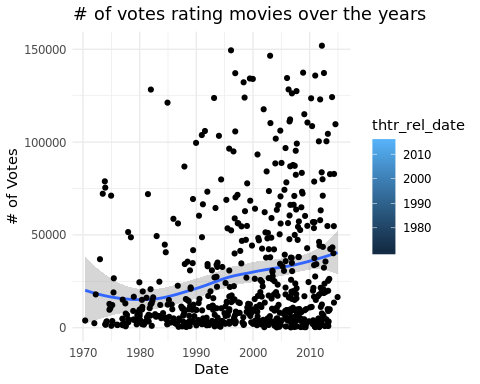
### - Do older films tend to have higher score/number of votes? if this is true, it could be because only good old films get remembered

Let’s see what is the amount of votes over the years that we have for all the films.

ggplot(movies, aes(fill = thtr\_rel\_date)) + geom\_smooth(aes(x = thtr\_rel\_date, y = imdb\_num\_votes),method = 'loess', formula = 'y ~ x') + geom\_point(aes(x = thtr\_rel\_date, y = imdb\_num\_votes)) + theme\_minimal() + xlab("Date") + ylab("# of Votes") + ylim(min(movies$imdb\_num\_votes),quantile(movies$imdb\_num\_votes,.90)) + ggtitle("# of votes rating movies over the years") # + scale\_fill\_discrete(name = "Decade", labels=c("70's","80's","90's","00's","10's"))

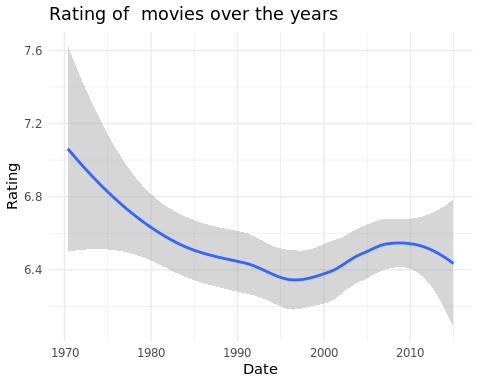
## Warning: Removed 65 rows containing non-finite values (stat\_smooth).

## Warning: Removed 65 rows containing missing values (geom\_point).



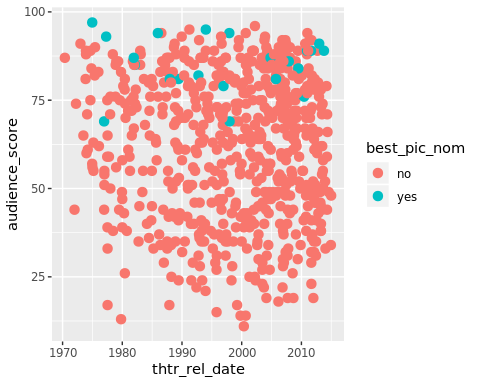
Okay so it seems like the number of votes does increase over the years, but does it afect the overall score?

ggplot(movies, aes(fill = thtr\_rel\_date)) + geom\_smooth(aes(x = thtr\_rel\_date, y = imdb\_rating), method = 'loess', formula = 'y ~ x'  
) + theme\_minimal() + xlab("Date") + ylab("Rating") + ggtitle("Rating of movies over the years") # + scale\_fill\_discrete(name = "Decade", labels=c("70's","80's","90's","00's","10's"))

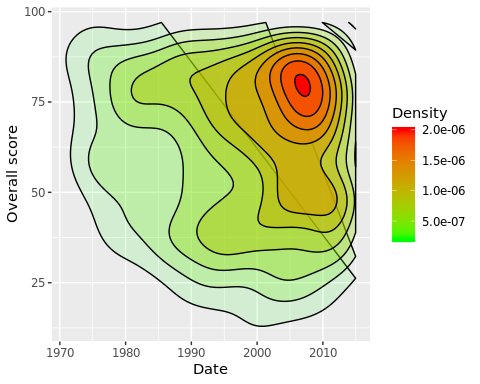


We can clearly see a decrease in overall score of this movies, but we can only asume that the newer the film, the more votes it gets and the worse it performs, but it could be because the quality in filmaking has woesened.

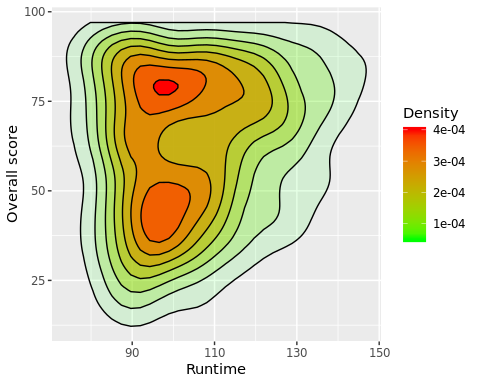
ggplot(data=movies, aes(x=thtr\_rel\_date, y=audience\_score)) + geom\_point(aes(col=best\_pic\_nom), size=3)



commonTheme = list(labs(color="Density",fill="Density", x="Date", y="Overall score"))  
ggplot(data=movies,aes(thtr\_rel\_date, audience\_score)) +   
 stat\_density2d(aes(fill=..level..,alpha=..level..),geom='polygon',colour='black') +   
 scale\_fill\_continuous(low="green",high="red") +  
 guides(alpha="none") +  
 commonTheme

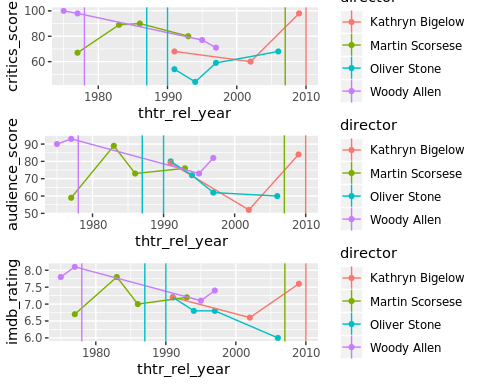


commonTheme = list(labs(color="Density",fill="Density", x="Runtime", y="Overall score"))  
ggplot(data=movies,aes(runtime, audience\_score)) +   
 stat\_density2d(aes(fill=..level..,alpha=..level..),geom='polygon',colour='black') +   
 scale\_fill\_continuous(low="green",high="red") +  
 guides(alpha="none") +  
 commonTheme



## 5. Do actors and directors have a higher score once they won? (find the first film and see the tendency, look for sudden change)

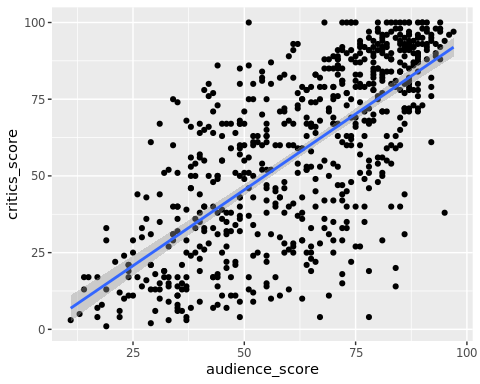
best\_dirs = subset(movies, movies$best\_dir\_win == "yes") %>%  
 group\_by(director) %>%  
 filter(n() > 2)  
  
win\_years = data.frame(  
 year = c(2010, 2007,1987, 1990, 1978),  
 director=c("Kathryn Bigelow", "Martin Scorsese","Oliver Stone","Oliver Stone","Woody Allen")  
)  
  
win\_year\_intersects =   
 geom\_vline(data = win\_years, aes(xintercept=year, color = director))  
  
p1 = ggplot(best\_dirs, aes(y=critics\_score, x=thtr\_rel\_year, color=director))+  
 geom\_point()+  
 geom\_line() +  
 win\_year\_intersects   
  
p2 = ggplot(best\_dirs, aes(y=audience\_score, x=thtr\_rel\_year, color=director))+  
 geom\_point()+  
 geom\_line() +  
 win\_year\_intersects  
  
p3= ggplot(best\_dirs, aes(y=imdb\_rating, x=thtr\_rel\_year, color=director))+  
 geom\_point()+  
 geom\_line() +  
 win\_year\_intersects  
  
grid.arrange(p1, p2, p3, nrow = 3)



#Provided questions ##1. What can we say about the relationship between audience scores and at least four of the other variables in this dataset?

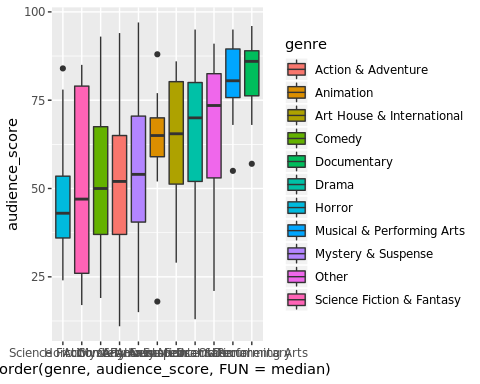
### Audience score vs Critics score

ggplot(movies, aes(x = audience\_score, y = critics\_score)) + geom\_point() + geom\_smooth(method =  
 lm)

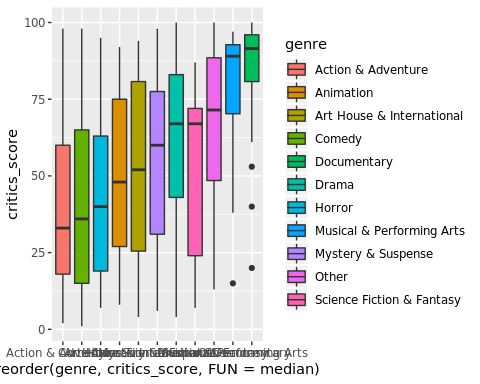


### Audience score vs Genre and Audience Score vs Critics Score vs Genre

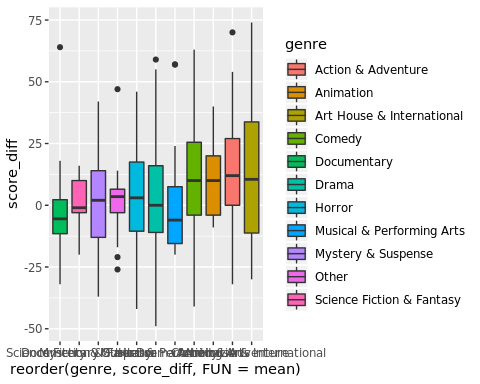
ggplot(movies, aes(  
 x = reorder(genre, audience\_score, FUN = median),  
 y = audience\_score,  
 fill = genre  
)) + geom\_boxplot()



ggplot(movies, aes(  
 x = reorder(genre, critics\_score, FUN = median),  
 y = critics\_score,  
 fill = genre  
)) + geom\_boxplot()



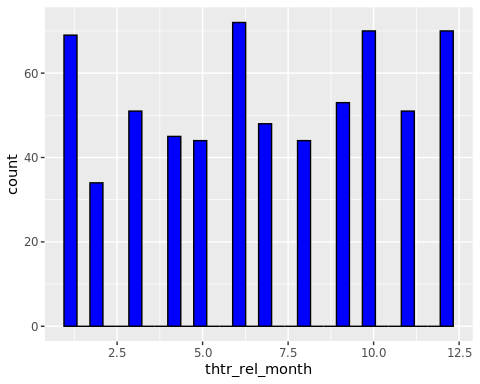
movies$score\_diff <- movies$audience\_score - movies$critics\_score  
ggplot(movies, aes(  
 x = reorder(genre, score\_diff, FUN = mean),  
 y = score\_diff,  
 fill = genre  
)) + geom\_boxplot()



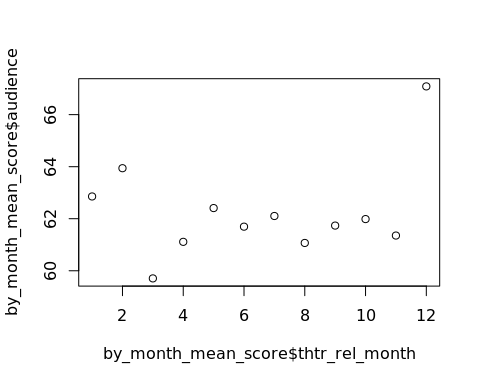
### Audience score vs Thtre\_rel\_month

ggplot(movies, aes(x = thtr\_rel\_month)) + geom\_histogram(color = "black", fill ="blue")

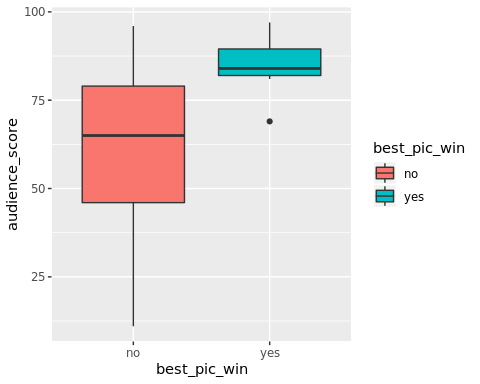
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



by\_month <- group\_by(movies, thtr\_rel\_month)  
by\_month\_mean\_score<-summarise(by\_month, audience=mean(audience\_score))  
plot(x=by\_month\_mean\_score$thtr\_rel\_month,y=by\_month\_mean\_score$audience)

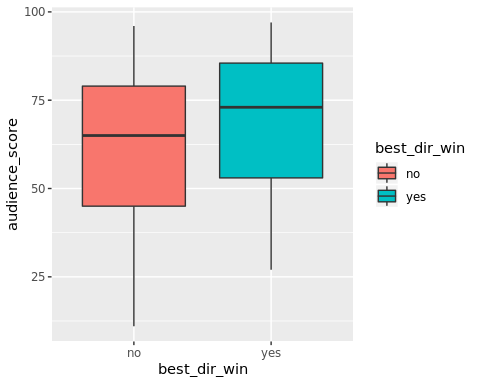
 ###Audience score vs Best\_pic\_win

ggplot(movies,  
 aes(x = best\_pic\_win, y = audience\_score, fill = best\_pic\_win)) + geom\_boxplot()



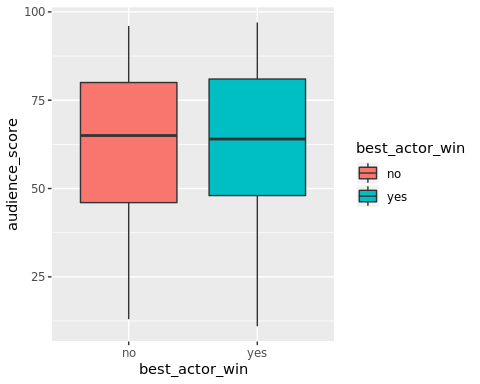
### Audience score vs Best\_dir\_win

ggplot(movies,  
 aes(x = best\_dir\_win, y = audience\_score, fill = best\_dir\_win)) + geom\_boxplot()



### Audience score vs Best\_actor\_win/Best\_actress\_win

ggplot(movies,  
 aes(x = best\_actor\_win, y = audience\_score, fill = best\_actor\_win)) +  
 geom\_boxplot()

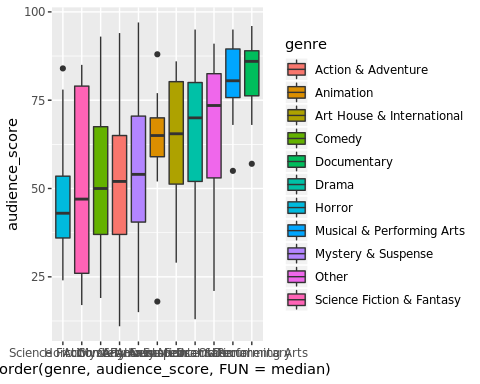


ggplot(movies,  
 aes(x = best\_actress\_win, y = audience\_score, fill = best\_actress\_win)) +  
 geom\_boxplot()

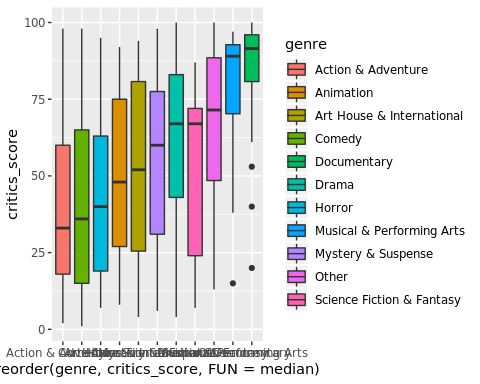


## 2. Is there a difference in the score given by audience members (or critics) for movies of different genres? And for movies with different MPAA ratings? If so, which categories seem to be significantly different for each other?

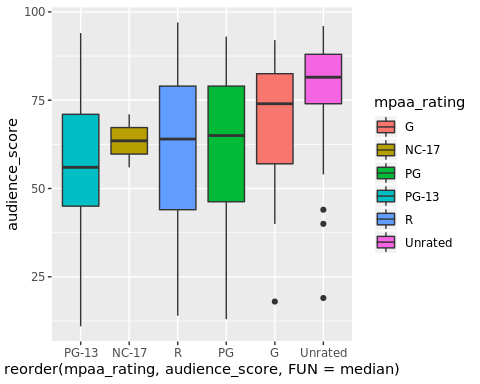
ggplot(movies, aes(  
 x = reorder(genre, audience\_score, FUN = median),  
 y = audience\_score,  
 fill = genre  
)) + geom\_boxplot()



ggplot(movies, aes(  
 x = reorder(genre, critics\_score, FUN = median),  
 y = critics\_score,  
 fill = genre  
)) + geom\_boxplot()



ggplot(movies, aes(  
 x = reorder(mpaa\_rating, audience\_score, FUN = median),  
 y = audience\_score,  
 fill = mpaa\_rating  
)) + geom\_boxplot()



ggplot(movies, aes(  
 x = reorder(mpaa\_rating, critics\_score, FUN = median),  
 y = critics\_score,  
 fill = mpaa\_rating  
)) + geom\_boxplot()

