Movie Correlation and Analysis in R

Ed Garcia

8/27/2021

Contents


```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

Project Questions

- In the past four decades, were high budgets, huge star power, or a franchise tag necessary to make a top grossing or top profitable movie?
- What other variables helped create a financially successful movie?
- What can be learned from movie trends of the last four decades?

Source Data and Inspiration

- This project used the "Movie Industry, Four Decades of Movies" data set posted on Kaggle by Daniel Grijalva.

- This project was inspired and informed from Alex Freberg's "Correlation in Python" tutorial project on YouTube.
- My mind works better in R than in Python, so I practiced translating Python code and analysis steps into R.
- I also translated my analysis into a user-friendly dashboard for public consumption. I was inspired and informed from the structure of Abhishek Agarrwal's "Tableau IMDB Movies Ratings Data Analysis and Dashboard Project Tutorial for Practice" project on YouTube.
- My published dashboard is available here.

PROJECT SETUP

Import libraries:

```
library(tidyverse) # I'm a big fan of dplyr
library(ggplot2) # for visualizations
library(reshape2) # for melting the correlation matrix
library(superml) # for encoding categorical data in correlation matrix
```

Import data: movies <- read.csv("filepath/filename.csv")

Glimpse the data:

```
glimpse(movies) # Released column should be formatted as a date
```

```
## Rows: 7,668
## Columns: 15
## $ name     <chr> "The Shining", "The Blue Lagoon", "Star Wars: Episode V - The~
## $ rating   <chr> "R", "R", "PG", "PG", "R", "R", "R", "R", "PG", "R", "PG", "P~
## $ genre    <chr> "Drama", "Adventure", "Action", "Comedy", "Comedy", "Horror",~
## $ year     <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1~
## $ released <chr> "June 13, 1980 (United States)", "July 2, 1980 (United States~
## $ score    <dbl> 8.4, 5.8, 8.7, 7.7, 7.3, 6.4, 7.9, 8.2, 6.8, 7.0, 6.1, 7.3, 5~
## $ votes    <dbl> 927000, 65000, 1200000, 221000, 108000, 123000, 188000, 33000~
## $ director <chr> "Stanley Kubrick", "Randal Kleiser", "Irvin Kershner", "Jim A~
## $ writer   <chr> "Stephen King", "Henry De Vere Stacpoole", "Leigh Brackett", ~
## $ star     <chr> "Jack Nicholson", "Brooke Shields", "Mark Hamill", "Robert Ha~
## $ country  <chr> "United Kingdom", "United States", "United States", "United S~
## $ budget   <dbl> 1.9e+07, 4.5e+06, 1.8e+07, 3.5e+06, 6.0e+06, 5.5e+05, 2.7e+07~
## $ gross    <dbl> 46998772, 58853106, 538375067, 83453539, 39846344, 39754601, ~
## $ company  <chr> "Warner Bros.", "Columbia Pictures", "Lucasfilm", "Paramount ~
## $ runtime  <dbl> 146, 104, 124, 88, 98, 95, 133, 129, 127, 100, 116, 109, 114,~
```

Summarize the data:

```
summary(movies) # nulls will need to be addressed
```

```
##      name              rating             genre                year
##  Length:7668        Length:7668        Length:7668        Min.   :1980
##  Class :character   Class :character   Class :character   1st Qu.:1991
##  Mode  :character   Mode  :character   Mode  :character   Median :2000
##                                                           Mean   :2000
##                                                           3rd Qu.:2010
##                                                           Max.   :2020
##
##    released             score          votes           director
```

```
##  Length:7668        Min.   :1.90   Min.   :      7   Length:7668
##  Class :character   1st Qu.:5.80   1st Qu.:   9100   Class :character
##  Mode  :character   Median :6.50   Median :  33000   Mode  :character
##                     Mean   :6.39   Mean   :  88109
##                     3rd Qu.:7.10   3rd Qu.:  93000
##                     Max.   :9.30   Max.   :2400000
##                     NA's   :3      NA's   :3
##     writer             star             country             budget
##  Length:7668        Length:7668        Length:7668        Min.   :     3000
##  Class :character   Class :character   Class :character   1st Qu.: 10000000
##  Mode  :character   Mode  :character   Mode  :character   Median : 20500000
##                                                           Mean   : 35589876
##                                                           3rd Qu.: 45000000
##                                                           Max.   :356000000
##                                                           NA's   :2171
##      gross              company             runtime
##  Min.   :3.090e+02   Length:7668        Min.   : 55.0
##  1st Qu.:4.532e+06   Class :character   1st Qu.: 95.0
##  Median :2.021e+07   Mode  :character   Median :104.0
##  Mean   :7.850e+07                      Mean   :107.3
##  3rd Qu.:7.602e+07                      3rd Qu.:116.0
##  Max.   :2.847e+09                      Max.   :366.0
##  NA's   :189                            NA's   :4
```

View the data:

```
head(movies, 10) # there are discrepancies between year and released columns
```

```
##                                                name rating     genre year
## 1                                       The Shining      R     Drama 1980
## 2                                     The Blue Lagoon      R Adventure 1980
## 3  Star Wars: Episode V - The Empire Strikes Back     PG    Action 1980
## 4                                         Airplane!     PG    Comedy 1980
## 5                                       Caddyshack      R    Comedy 1980
## 6                                   Friday the 13th      R    Horror 1980
## 7                                The Blues Brothers      R    Action 1980
## 8                                       Raging Bull      R Biography 1980
## 9                                       Superman II     PG    Action 1980
## 10                                  The Long Riders      R Biography 1980
##                              released score   votes           director
## 1      June 13, 1980 (United States)   8.4  927000    Stanley Kubrick
## 2       July 2, 1980 (United States)   5.8   65000     Randal Kleiser
## 3      June 20, 1980 (United States)   8.7 1200000     Irvin Kershner
## 4       July 2, 1980 (United States)   7.7  221000       Jim Abrahams
## 5      July 25, 1980 (United States)   7.3  108000       Harold Ramis
## 6        May 9, 1980 (United States)   6.4  123000 Sean S. Cunningham
## 7      June 20, 1980 (United States)   7.9  188000        John Landis
## 8  December 19, 1980 (United States)   8.2  330000    Martin Scorsese
## 9      June 19, 1981 (United States)   6.8  101000     Richard Lester
## 10      May 16, 1980 (United States)   7.0   10000        Walter Hill
##                      writer            star        country  budget     gross
## 1              Stephen King  Jack Nicholson United Kingdom 1.9e+07  46998772
## 2  Henry De Vere Stacpoole  Brooke Shields  United States 4.5e+06  58853106
## 3           Leigh Brackett     Mark Hamill  United States 1.8e+07 538375067
## 4             Jim Abrahams     Robert Hays  United States 3.5e+06  83453539
```

```
## 5       Brian Doyle-Murray     Chevy Chase  United States 6.0e+06  39846344
## 6            Victor Miller    Betsy Palmer  United States 5.5e+05  39754601
## 7              Dan Aykroyd    John Belushi  United States 2.7e+07 115229890
## 8             Jake LaMotta  Robert De Niro  United States 1.8e+07  23402427
## 9             Jerry Siegel    Gene Hackman  United States 5.4e+07 108185706
## 10             Bill Bryden David Carradine  United States 1.0e+07  15795189
##                         company runtime
## 1                   Warner Bros.     146
## 2              Columbia Pictures     104
## 3                       Lucasfilm     124
## 4             Paramount Pictures      88
## 5                 Orion Pictures      98
## 6             Paramount Pictures      95
## 7             Universal Pictures     133
## 8  Chartoff-Winkler Productions     129
## 9                 Dovemead Films     127
## 10                United Artists     100
```

DATA CLEANING

Discrepancy between the Year and Released columns: they do not always align.

After researching the source data on IMDB, I determined that:

- the Year column refers to the year of the movie's first premiere showing
- the Released column refers to the date of the movie's first full-scale release
- The country in which the movie was released is parenthesized in the Release column

Clean the Year and Released columns to clarify the discrepancy.

Rename Year column to Premiere Year:

```
movies <- movies %>%
  rename(premiere = year)
```

Split Released column into Full Release Date and Full Release Location columns:

```
# this creates the separated date and location
split <- unlist(strsplit(movies$released, split="[()]"))
# this creates the new column names
cols <- c("full_release_date", "full_release_location")
# the following steps set up the re-insertion into the data frame
nC <- length(cols)
ind <- seq(from=1, by=nC, length=nrow(movies))
for(i in 1:nC) {movies[, cols[i]] <- split[ind + i - 1]}
```

Change Full Release Date column to date format:

```
movies$full_release_date <- as.Date(movies$full_release_date, "%B %d, %Y")
```

Remove the now-irrelevant Released column:

```
movies <- movies %>%
  select(-c(released))
```

Drop any duplicate rows:

```
nrow(movies) # 7668 total rows
```

```
## [1] 7668
```

```
nrow(distinct(movies)) # 7668 total distinct rows
```

```
## [1] 7668
```

```
# there are no duplicate rows
```

Determine profit for each movie.

Make a Profit column:

```
movies$profit <- movies$gross - movies$budget
```

Convert the gross, budget, and profit to millions to increase readability:

```
movies$gross <- movies$gross / 1000000
movies$budget <- movies$budget / 1000000
movies$profit <- movies$profit / 1000000
movies <- movies %>%
rename(grossM = gross, budgetM = budget, profitM = profit)
```

Make a Profit Percentage column:

```
movies$profit_percent <- (movies$profitM / movies$budgetM) * 100
```

Clean the Rating column.

Find out what ratings are used in this data set:

```
unique(movies$rating)
```

```
##  [1] "R"         "PG"        "G"         ""          "Not Rated" "NC-17"
##  [7] "Approved"  "TV-PG"     "PG-13"     "Unrated"   "X"         "TV-MA"
## [13] "TV-14"
```

Unite Unrated and Not Rated ratings:

```
movies["rating"][movies["rating"] == "Not Rated"] <- "Unrated"
```

Print the movies that are rated Approved:

```
subset(movies, rating == "Approved")
```

```
##                   name   rating     genre premiere score votes   director
## 121 Tarzan the Ape Man Approved Adventure     1981   3.4  5300 John Derek
##       writer     star       country budgetM   grossM                  company
## 121 Tom Rowe Bo Derek United States     6.5 36.56528 Metro-Goldwyn-Mayer (MGM)
##     runtime full_release_date full_release_location  profitM profit_percent
## 121     115        1981-07-24         United States 30.06528       462.5428
```

```
# The only movie with an Approved rating is "Tarzan the Ape Man".
  # Upon further research on IMDB, the movie poster states that the movie is rated R.
```

Unite Approved and R ratings:

```
movies["rating"][movies["rating"] == "Approved"] <- "R"
```

Unite other ratings as Other:

```
movies["rating"][movies["rating"] == ""] <- "Other"
movies["rating"][movies["rating"] == "X"] <- "Other"
movies["rating"][movies["rating"] == "NC-17"] <- "Other"
movies["rating"][movies["rating"] == "TV-PG"] <- "Other"
movies["rating"][movies["rating"] == "TV-14"] <- "Other"
movies["rating"][movies["rating"] == "TV-MA"] <- "Other"
```

Clean the Genre column, create a Decades column, and a note about NULLs.

Find out what genres are used in this data set:

```
unique(movies$genre)
```

```
##  [1] "Drama"     "Adventure" "Action"    "Comedy"    "Horror"    "Biography"
##  [7] "Crime"     "Fantasy"   "Family"    "Sci-Fi"    "Animation" "Romance"
## [13] "Music"     "Western"   "Thriller"  "History"   "Mystery"   "Sport"
## [19] "Musical"
```

Unite Musical and Music ratings:

```
movies["genre"][movies["genre"] == "Music"] <- "Musical" # Only one entry as Music
movies["genre"][movies["genre"] == "History"] <- "Biography" # Only one entry as History
movies["genre"][movies["genre"] == "Sport"] <- "Drama" # Only one entry as Sport
```

Create a decades column:

```
movies$decade <- as.numeric(format(movies$full_release_date, format="%Y"))
movies$decade <- round(movies$decade, -1)
```

I am leaving null values in the data set, and will remove them on a case-by-case scenario.

DATA EXPLORATION

Which variables (stars, directors, writers, companies) were involved in the highest grossing movies?

```
movies %>%
  select(c(name, premiere, star, director, writer, company,
           profitM, profit_percent, grossM)) %>%
  arrange(desc(grossM)) %>%
  top_n(20)
```

```
##                                          name premiere              star
## 1                                      Avatar     2009   Sam Worthington
## 2                           Avengers: Endgame     2019 Robert Downey Jr.
## 3                                     Titanic     1997 Leonardo DiCaprio
## 4  Star Wars: Episode VII - The Force Awakens     2015      Daisy Ridley
## 5                    Avengers: Infinity War     2018 Robert Downey Jr.
## 6                             The Lion King     2019     Donald Glover
## 7                           Jurassic World     2015       Chris Pratt
## 8                              The Avengers     2012 Robert Downey Jr.
## 9                                 Furious 7     2015        Vin Diesel
## 10                                Frozen II     2019      Kristen Bell
## 11                  Avengers: Age of Ultron     2015 Robert Downey Jr.
## 12                            Black Panther     2018  Chadwick Boseman
## 13 Harry Potter and the Deathly Hallows: Part 2     2011  Daniel Radcliffe
## 14      Star Wars: Episode VIII - The Last Jedi     2017      Daisy Ridley
```

```
## 15               Jurassic World: Fallen Kingdom     2018       Chris Pratt
## 16                                       Frozen     2013      Kristen Bell
## 17                         Beauty and the Beast     2017       Emma Watson
## 18                                Incredibles 2     2018   Craig T. Nelson
## 19                       The Fate of the Furious     2017        Vin Diesel
## 20                                   Iron Man 3     2013 Robert Downey Jr.
##           director              writer                        company   profitM
## 1    James Cameron       James Cameron         Twentieth Century Fox 2610.2462
## 2    Anthony Russo  Christopher Markus                Marvel Studios 2441.5013
## 3    James Cameron       James Cameron         Twentieth Century Fox 2001.6473
## 4      J.J. Abrams     Lawrence Kasdan                     Lucasfilm 1824.5217
## 5    Anthony Russo  Christopher Markus                Marvel Studios 1727.3598
## 6      Jon Favreau      Jeff Nathanson          Walt Disney Pictures 1410.7276
## 7  Colin Trevorrow          Rick Jaffa             Universal Pictures 1520.5164
## 8      Joss Whedon         Joss Whedon                Marvel Studios 1298.8155
## 9        James Wan        Chris Morgan             Universal Pictures 1325.3414
## 10      Chris Buck        Jennifer Lee Walt Disney Animation Studios 1300.0269
## 11     Joss Whedon         Joss Whedon                Marvel Studios 1152.8095
## 12    Ryan Coogler        Ryan Coogler                Marvel Studios 1147.5980
## 13     David Yates        Steve Kloves                  Warner Bros. 1217.3217
## 14    Rian Johnson        Rian Johnson          Walt Disney Pictures 1015.6988
## 15     J.A. Bayona      Derek Connolly             Universal Pictures 1140.4663
## 16      Chris Buck        Jennifer Lee Walt Disney Animation Studios 1131.5081
## 17     Bill Condon     Stephen Chbosky               Mandeville Films 1104.4345
## 18       Brad Bird           Brad Bird          Walt Disney Pictures 1044.6395
## 19    F. Gary Gray Gary Scott Thompson             Universal Pictures  986.0051
## 20     Shane Black         Drew Pearce                Marvel Studios 1014.8113
##    profit_percent   grossM
## 1       1101.3697 2847.246
## 2        685.8150 2797.501
## 3       1000.8236 2201.647
## 4        744.7027 2069.522
## 5        538.1183 2048.360
## 6        542.5875 1670.728
## 7       1013.6776 1670.516
## 8        590.3707 1518.816
## 9        697.5481 1515.341
## 10       866.6846 1450.027
## 11       461.1238 1402.810
## 12       573.7990 1347.598
## 13       973.8573 1342.322
## 14       320.4097 1332.699
## 15       670.8625 1310.466
## 16       754.3387 1281.508
## 17       690.2716 1264.435
## 18       522.3198 1244.640
## 19       394.4020 1236.005
## 20       507.4056 1214.811
```

Except for Avatar and Titanic, all the Top 20 grossing movies premiered in the last decade.

- This makes sense, since movies' gross will continue to rise due to inflation's impact on the cost of movie ticket sales.

There were many stars that appear multiple times.

- This can be explained by the fact that many of these movies are franchises (Avengers, Star Wars, Furious) which utilize multi-movie contracts with their stars

Other highlights:

- James Cameron appears twice in the top 3 of director and writer
- No female directors. Jennifer Lee was the only female writer, appearing twice for the Frozen movies
- Only 7 companies: 20th Century Fox, Marvel, Lucasfilm, Disney, Universal, Warner Bros, Mandeville

Which variables (stars, directors, writers, companies) were involved in the most profitable ($) movies?

```
movies %>%
  select(c(name, premiere, star, director, writer, company, profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(20)
```

```
##                                          name premiere              star
## 1                                      Avatar     2009   Sam Worthington
## 2                           Avengers: Endgame     2019 Robert Downey Jr.
## 3                                     Titanic     1997 Leonardo DiCaprio
## 4    Star Wars: Episode VII - The Force Awakens     2015      Daisy Ridley
## 5                       Avengers: Infinity War     2018 Robert Downey Jr.
## 6                              Jurassic World     2015       Chris Pratt
## 7                                The Lion King     2019     Donald Glover
## 8                                    Furious 7     2015        Vin Diesel
## 9                                    Frozen II     2019      Kristen Bell
## 10                                The Avengers     2012 Robert Downey Jr.
## 11 Harry Potter and the Deathly Hallows: Part 2     2011  Daniel Radcliffe
## 12                      Avengers: Age of Ultron     2015 Robert Downey Jr.
## 13                                Black Panther     2018  Chadwick Boseman
## 14               Jurassic World: Fallen Kingdom     2018       Chris Pratt
## 15                                       Frozen     2013      Kristen Bell
## 16                         Beauty and the Beast     2017       Emma Watson
## 17                                     Minions     2015    Sandra Bullock
## 18 The Lord of the Rings: The Return of the King     2003       Elijah Wood
## 19                                Incredibles 2     2018   Craig T. Nelson
## 20                                The Lion King     1994 Matthew Broderick
##           director             writer                        company  profitM
## 1    James Cameron      James Cameron         Twentieth Century Fox 2610.246
## 2    Anthony Russo Christopher Markus                Marvel Studios 2441.501
## 3    James Cameron      James Cameron         Twentieth Century Fox 2001.647
## 4      J.J. Abrams    Lawrence Kasdan                     Lucasfilm 1824.522
## 5    Anthony Russo Christopher Markus                Marvel Studios 1727.360
## 6  Colin Trevorrow         Rick Jaffa            Universal Pictures 1520.516
## 7      Jon Favreau     Jeff Nathanson          Walt Disney Pictures 1410.728
## 8        James Wan       Chris Morgan            Universal Pictures 1325.341
## 9       Chris Buck       Jennifer Lee Walt Disney Animation Studios 1300.027
## 10     Joss Whedon        Joss Whedon                Marvel Studios 1298.816
## 11     David Yates       Steve Kloves                  Warner Bros. 1217.322
## 12     Joss Whedon        Joss Whedon                Marvel Studios 1152.810
## 13    Ryan Coogler       Ryan Coogler                Marvel Studios 1147.598
## 14     J.A. Bayona     Derek Connolly            Universal Pictures 1140.466
## 15      Chris Buck       Jennifer Lee Walt Disney Animation Studios 1131.508
## 16     Bill Condon    Stephen Chbosky              Mandeville Films 1104.435
```

```
## 17      Kyle Balda        Brian Lynch    Illumination Entertainment 1085.445
## 18   Peter Jackson     J.R.R. Tolkien               New Line Cinema 1052.031
## 19       Brad Bird          Brad Bird          Walt Disney Pictures 1044.640
## 20    Roger Allers       Irene Mecchi          Walt Disney Pictures 1038.721
```

Not much difference between the variables in the Top 20 profitable ($) list and the Top 20 grossing list (this was expected).

Which variables (stars, directors, writers, companies) were involved in the movies with the highest profit percentage?

```
movies %>%
  select(c(name, premiere, star, director, writer, company, grossM, profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(20)
```

```
##                          name premiere              star           director
## 1         Paranormal Activity     2007 Katie Featherston          Oren Peli
## 2     The Blair Witch Project     1999   Heather Donahue      Daniel Myrick
## 3                 The Gallows     2015     Reese Mishler       Travis Cluff
## 4                 El Mariachi     1992   Carlos Gallardo   Robert Rodriguez
## 5                        Once     2007      Glen Hansard        John Carney
## 6                      Clerks     1994  Brian O'Halloran        Kevin Smith
## 7            Napoleon Dynamite     2004         Jon Heder         Jared Hess
## 8       In the Company of Men     1997     Aaron Eckhart        Neil LaBute
## 9                 Keeping Mum     2005    Rowan Atkinson      Niall Johnson
## 10                Open Water     2003    Blanchard Ryan       Chris Kentis
## 11           The Devil Inside     2012  Fernanda Andrade William Brent Bell
## 12             The Quiet Ones     2014      Jared Harris         John Pogue
## 13                        Saw     2004        Cary Elwes          James Wan
## 14                  Searching     2018          John Cho    Aneesh Chaganty
## 15                     Primer     2004     Shane Carruth      Shane Carruth
## 16 E.T. the Extra-Terrestrial     1982      Henry Thomas   Steven Spielberg
## 17   My Big Fat Greek Wedding     2002      Nia Vardalos         Joel Zwick
## 18             The Full Monty     1997    Robert Carlyle     Peter Cattaneo
## 19            Friday the 13th     1980      Betsy Palmer Sean S. Cunningham
## 20                  Fireproof     2008      Kirk Cameron      Alex Kendrick
##                writer                        company     grossM
## 1           Oren Peli                    Solana Films 193.355800
## 2       Daniel Myrick                     Haxan Films 248.639099
## 3        Chris Lofing                 New Line Cinema  42.964410
## 4    Robert Rodriguez                Columbia Pictures   2.040920
## 5         John Carney      BÃ³rd ScannÃ¡n na hÃ‰ireann  20.936722
## 6         Kevin Smith           View Askew Productions   3.151130
## 7          Jared Hess         Fox Searchlight Pictures  46.138887
## 8         Neil LaBute Alliance Atlantis Communications   2.804473
## 9       Richard Russo              Summit Entertainment  18.586834
## 10       Chris Kentis              Plunge Pictures LLC  54.683487
## 11 William Brent Bell                 Insurge Pictures 101.758490
## 12    Craig Rosenberg            Exclusive Media Group  17.835162
## 13     Leigh Whannell          Evolution Entertainment 103.911669
## 14    Aneesh Chaganty                     Screen Gems  75.462037
## 15      Shane Carruth                            ERBP   0.545436
## 16   Melissa Mathison               Universal Pictures 792.910554
## 17       Nia Vardalos                Gold Circle Films 368.744044
## 18      Simon Beaufoy                    Redwave Films 257.938649
```

```
## 19      Victor Miller              Paramount Pictures  39.754601
## 20      Alex Kendrick            Samuel Goldwyn Films  33.473297
##    profit_percent
## 1     1288938.667
## 2      414298.498
## 3       42864.410
## 4       29056.000
## 5       13857.815
## 6       11570.852
## 7       11434.722
## 8       11117.892
## 9       10898.127
## 10      10836.697
## 11      10075.849
## 12       8817.581
## 13       8559.306
## 14       8475.231
## 15       7691.943
## 16       7451.529
## 17       7274.881
## 18       7269.676
## 19       7128.109
## 20       6594.659
```

The Top 20 profitable (%) list contains more variety compared to the Top 20 grossing and Top 20 profitable ($) lists.

- There was at least 1 movie in each decade (80s, 90s, 00s, 10s).
- No sequels (i.e., these movies were not franchises when they were created).
- There were no stars, directors, writers, or companies that appear more than once.
- Only 5 movies had budgets of >= $1M, and none are in the top 10.
- In general, these movies succeeded in profitability despite their low budget.

Insights:

The highest profit (%) in the top 20 grossing movies is 1101% by Avatar.

- This is far lower than the profit (%) in the top 20 profit (%) movies.
- The lowest profit (%) in the profit (%) list is Fireproof with 6594%

None of the movies in the top 20 profit (%) movies had a gross above $370M.

- Except E.T. which had a gross of $792M

All of the movies in the top 20 gross movies had a profit ($) above $1B.

- except Fate of the Furious which had a profit ($) of $986M

CORRELATION MATRIX

Make correlation matrix for all variables:

```
labmovies <- movies # separate data frame for labels
label <- LabelEncoder$new()
# non-numerical variables are converted through label encoding:
labmovies$name <- label$fit_transform(labmovies$name)
```

```
labmovies$rating <- label$fit_transform(labmovies$rating)
labmovies$genre <- label$fit_transform(labmovies$genre)
labmovies$director <- label$fit_transform(labmovies$director)
labmovies$writer <- label$fit_transform(labmovies$writer)
labmovies$star <- label$fit_transform(labmovies$star)
labmovies$country <- label$fit_transform(labmovies$country)
labmovies$company <- label$fit_transform(labmovies$company)
labmovies$full_release_location <- label$fit_transform(labmovies$full_release_location)
head(labmovies) # check that numeric lables were applied correctly
```

```
##   name rating genre premiere score   votes director writer star country budgetM
## 1    0      0     0     1980   8.4  927000        0      0    0       0   19.00
## 2    1      0     1     1980   5.8   65000        1      1    1       1    4.50
## 3    2      1     2     1980   8.7 1200000        2      2    2       1   18.00
## 4    3      1     3     1980   7.7  221000        3      3    3       1    3.50
## 5    4      0     3     1980   7.3  108000        4      4    4       1    6.00
## 6    5      0     4     1980   6.4  123000        5      5    5       1    0.55
##      grossM company runtime full_release_date full_release_location   profitM
## 1  46.99877       0     146        1980-06-13                     0  27.99877
## 2  58.85311       1     104        1980-07-02                     0  54.35311
## 3 538.37507       2     124        1980-06-20                     0 520.37507
## 4  83.45354       3      88        1980-07-02                     0  79.95354
## 5  39.84634       4      98        1980-07-25                     0  33.84634
## 6  39.75460       3      95        1980-05-09                     0  39.20460
##   profit_percent decade
## 1       147.3620   1980
## 2      1207.8468   1980
## 3      2890.9726   1980
## 4      2284.3868   1980
## 5       564.1057   1980
## 6      7128.1093   1980
```

```
corr_matrix <- round(cor(labmovies[, sapply(labmovies, is.numeric)],
                         use = "complete.obs", method = "pearson"), 2)
```

Create a refined correlation heat map.

Get lower triangle of the correlation matrix:

```
get_lower_tri<-function(corr_matrix)
  {
  corr_matrix[upper.tri(corr_matrix)] <- NA
  return(corr_matrix)
  }
```

Get upper triangle of the correlation matrix:

```
get_upper_tri <- function(corr_matrix)
  {
  corr_matrix[lower.tri(corr_matrix)]<- NA
  return(corr_matrix)
  }
```

Return usable data frame:

```
upper_tri <- get_upper_tri(corr_matrix)
upper_tri
```

```
##                       name rating genre premiere score votes director writer
## name                     1   0.16  0.05     0.95  0.05  0.19     0.70   0.76
## rating                  NA   1.00 -0.10     0.18 -0.07  0.09     0.09   0.12
## genre                   NA     NA  1.00     0.05  0.05  0.01     0.06   0.05
## premiere                NA     NA    NA     1.00  0.05  0.20     0.73   0.78
## score                   NA     NA    NA       NA  1.00  0.47     0.00   0.02
## votes                   NA     NA    NA       NA    NA  1.00     0.09   0.11
## director                NA     NA    NA       NA    NA    NA     1.00   0.69
## writer                  NA     NA    NA       NA    NA    NA       NA   1.00
## star                    NA     NA    NA       NA    NA    NA       NA     NA
## country                 NA     NA    NA       NA    NA    NA       NA     NA
## budgetM                 NA     NA    NA       NA    NA    NA       NA     NA
## grossM                  NA     NA    NA       NA    NA    NA       NA     NA
## company                 NA     NA    NA       NA    NA    NA       NA     NA
## runtime                 NA     NA    NA       NA    NA    NA       NA     NA
## full_release_location   NA     NA    NA       NA    NA    NA       NA     NA
## profitM                 NA     NA    NA       NA    NA    NA       NA     NA
## profit_percent          NA     NA    NA       NA    NA    NA       NA     NA
## decade                  NA     NA    NA       NA    NA    NA       NA     NA
##                       star country budgetM grossM company runtime
## name                  0.68    0.09    0.30   0.24    0.50    0.06
## rating                0.12    0.00    0.26   0.21   -0.04    0.05
## genre                 0.06    0.00    0.07   0.09    0.02   -0.18
## premiere              0.71    0.09    0.33   0.27    0.51    0.07
## score                 0.00    0.08    0.07   0.22    0.04    0.42
## votes                 0.09   -0.02    0.44   0.61   -0.05    0.35
## director              0.62    0.07    0.09   0.14    0.45   -0.13
## writer                0.60    0.09    0.18   0.15    0.44   -0.01
## star                  1.00    0.09    0.11   0.14    0.41   -0.05
## country                 NA    1.00   -0.03  -0.04    0.13    0.08
## budgetM                 NA      NA    1.00   0.74   -0.15    0.32
## grossM                  NA      NA      NA   1.00   -0.09    0.28
## company                 NA      NA      NA     NA    1.00   -0.05
## runtime                 NA      NA      NA     NA      NA    1.00
## full_release_location   NA      NA      NA     NA      NA      NA
## profitM                 NA      NA      NA     NA      NA      NA
## profit_percent          NA      NA      NA     NA      NA      NA
## decade                  NA      NA      NA     NA      NA      NA
##                       full_release_location profitM profit_percent decade
## name                                   0.15    0.21           0.01   0.92
## rating                                 0.00    0.17          -0.02   0.16
## genre                                  0.01    0.09           0.01   0.05
## premiere                               0.15    0.24           0.01   0.97
## score                                  0.00    0.24           0.00   0.06
## votes                                 -0.05    0.61           0.02   0.18
## director                               0.13    0.14           0.02   0.71
## writer                                 0.13    0.13           0.02   0.76
## star                                   0.11    0.13           0.02   0.69
## country                                0.06   -0.03           0.00   0.09
## budgetM                               -0.06    0.61          -0.02   0.31
## grossM                                -0.05    0.98           0.02   0.26
## company                                0.16   -0.07           0.02   0.51
## runtime                               -0.04    0.24          -0.02   0.07
## full_release_location                  1.00   -0.05           0.00   0.13
```

```
## profitM                                  NA    1.00           0.02   0.22
## profit_percent                           NA      NA           1.00   0.01
## decade                                   NA      NA             NA   1.00
```

Helper function to reorder the correlation matrix :

```
reorder_corr_matrix <- function(corr_matrix)
  {
  # Use correlation between variables as distance
  dd <- as.dist((1-corr_matrix)/2)
  hc <- hclust(dd)  # hc = hierarchical clustering
  corr_matrix <-corr_matrix[hc$order, hc$order]
  }
```

Reorder the correlation matrix:

```
corr_matrix <- reorder_corr_matrix(corr_matrix)
upper_tri <- get_upper_tri(corr_matrix)
```

Melt the correlation matrix for plotting:

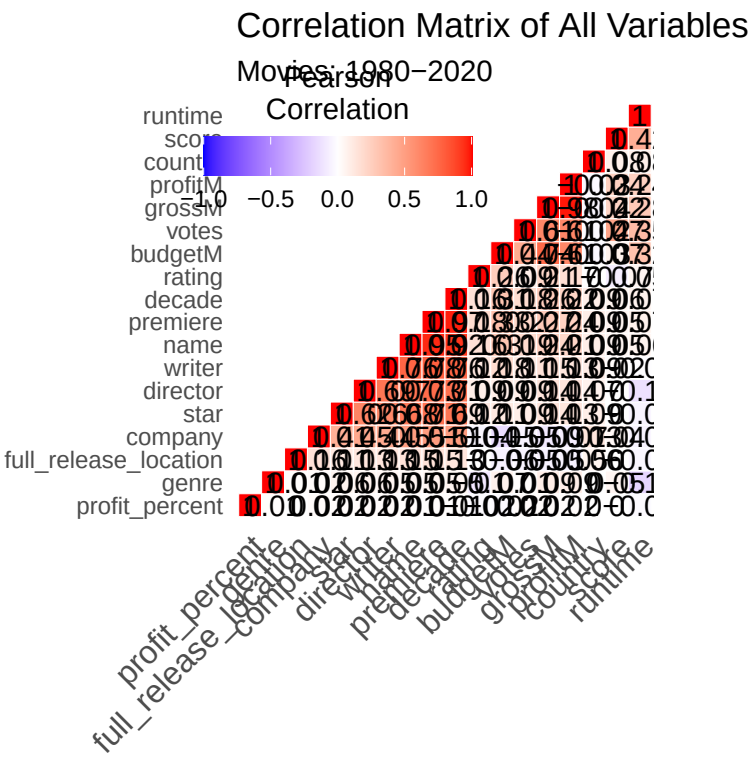
```
melted_corr_matrix <- melt(upper_tri, na.rm = TRUE)
```

Plot the heat map:

```
ggheatmap <- ggplot(data = melted_corr_matrix, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()
```

Add labels and text to plot:

```
ggheatmap +
  labs(x = "Movie Features", y = "Movie Features",
       title = "Correlation Matrix of All Variables",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```



Source:
'Movie Industry, Four Decades of Movies' IMDB dataset,
posted on Kaggle by Daniel Grijalva

These correlations are highly logical, such as year-related variables (decades and premieres), and collaboration-related variables (directors and writers often pair up together multiple times, as do stars).

Country was not highly correlated to other variables, nor was runtime, genre, rating, or profit percentage.

Since I am mostly interested in correlations with gross and profit, I will inspect this more closely.

CORRELATION LISTS AND SCATTER PLOTS OF SELECTED VARIABLES

Make a list of variables that were highly correlated to gross:

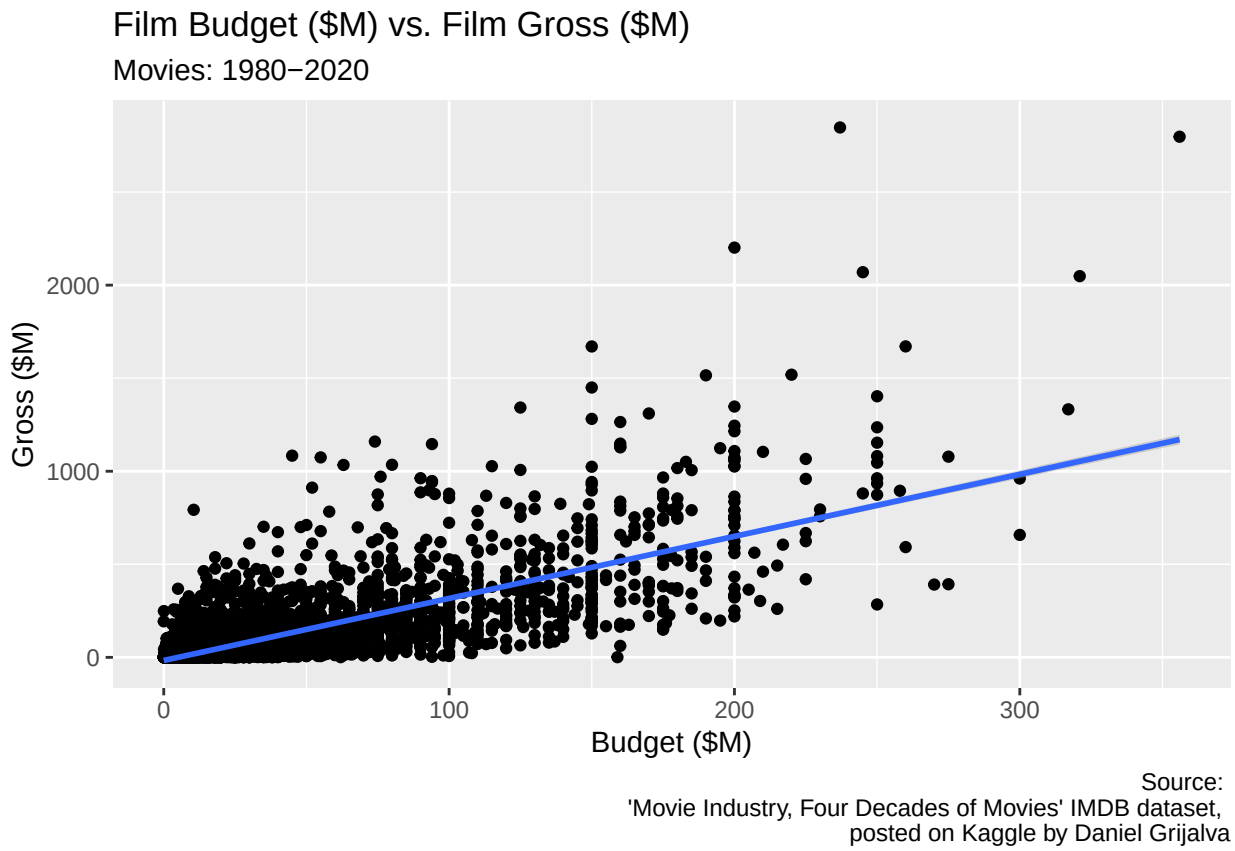
```
gross_high_corr <- melted_corr_matrix %>%
  filter(melted_corr_matrix$Var2 == "grossM"
         & melted_corr_matrix$value >= 0.5 # this only pulls highly correlated variables
         & melted_corr_matrix$value != 1) %>% # this ignores self-correlated variables
  arrange(desc(value))
tibble(gross_high_corr) # budget and votes have the highest correlation to gross
```

```
## # A tibble: 2 x 3
##   Var1    Var2   value
##   <fct>   <fct>  <dbl>
## 1 budgetM grossM  0.74
## 2 votes   grossM  0.61
```

As a reminder, Votes refers to the number of votes that the movies has obtained from IMDB users.

Create a scatter plot with budget vs gross:

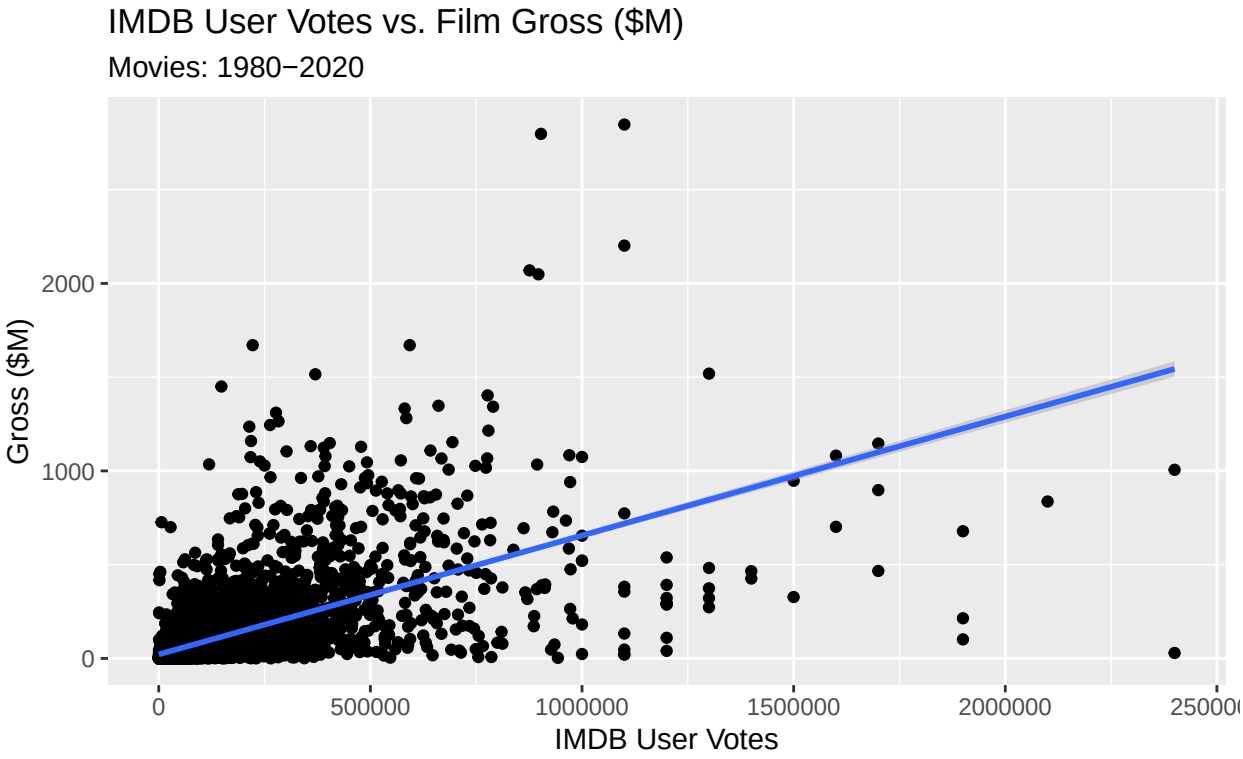
```
# the plot will skip over any NULLs
ggplot(movies, aes(budgetM, grossM)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Budget ($M)", y = "Gross ($M)",
       title = "Film Budget ($M) vs. Film Gross ($M)",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



Insights: A higher budget helped provide the potential for a higher gross.

Create a scatter plot with votes vs gross:

```
ggplot(movies, aes(votes, grossM)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "IMDB User Votes", y = "Gross ($M)",
       title = "IMDB User Votes vs. Film Gross ($M)",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



Insights: There were many votes for movies without a high gross. This is possibly because less financially successful movies can gain "cult" followings by certain demographics

Make a list of variables that were highly correlated to profit ($M):

```
profit_high_corr <- melted_corr_matrix %>%
  filter(melted_corr_matrix$Var2 == "profitM"
         & melted_corr_matrix$value >= 0.5
         & melted_corr_matrix$value != 1) %>%
  arrange(desc(value))
tibble(profit_high_corr)
```

```
## # A tibble: 3 x 3
##   Var1    Var2    value
##   <fct>   <fct>   <dbl>
## 1 grossM  profitM  0.98
## 2 budgetM profitM  0.61
## 3 votes   profitM  0.61
```

As expected, profit ($M) was correlated to both gross and budget (and votes).

Make a list of variables that were highly correlated to profit (%):

```
profperc_high_corr <- melted_corr_matrix %>%
  filter(melted_corr_matrix$Var2 == "profit_percent"
         & melted_corr_matrix$value >= 0.5
         & melted_corr_matrix$value != 1) %>%
  arrange(desc(value))
tibble(profperc_high_corr)
```

```
## # A tibble: 0 x 3
## # ... with 3 variables: Var1 <fct>, Var2 <fct>, value <dbl>
```

No high correlations found for profit (%). No scatter plot necessary.

Correlation Insights: The most profitable (%) movies did not have the highest budgets or gross, but the larger budgets tended to create larger gross.

FURTHER ANALYSIS BY CATEGORY (TOTAL RANGE, TOP GROSSING, TOP PROFITABLE($, %), TOP DECADE)

Here I will continue to investigate variables involved in top grossing and top profitable movies.

I will also analyze variable involvement throughout the total range of the data set, and variable involvement divided into decade ranges.

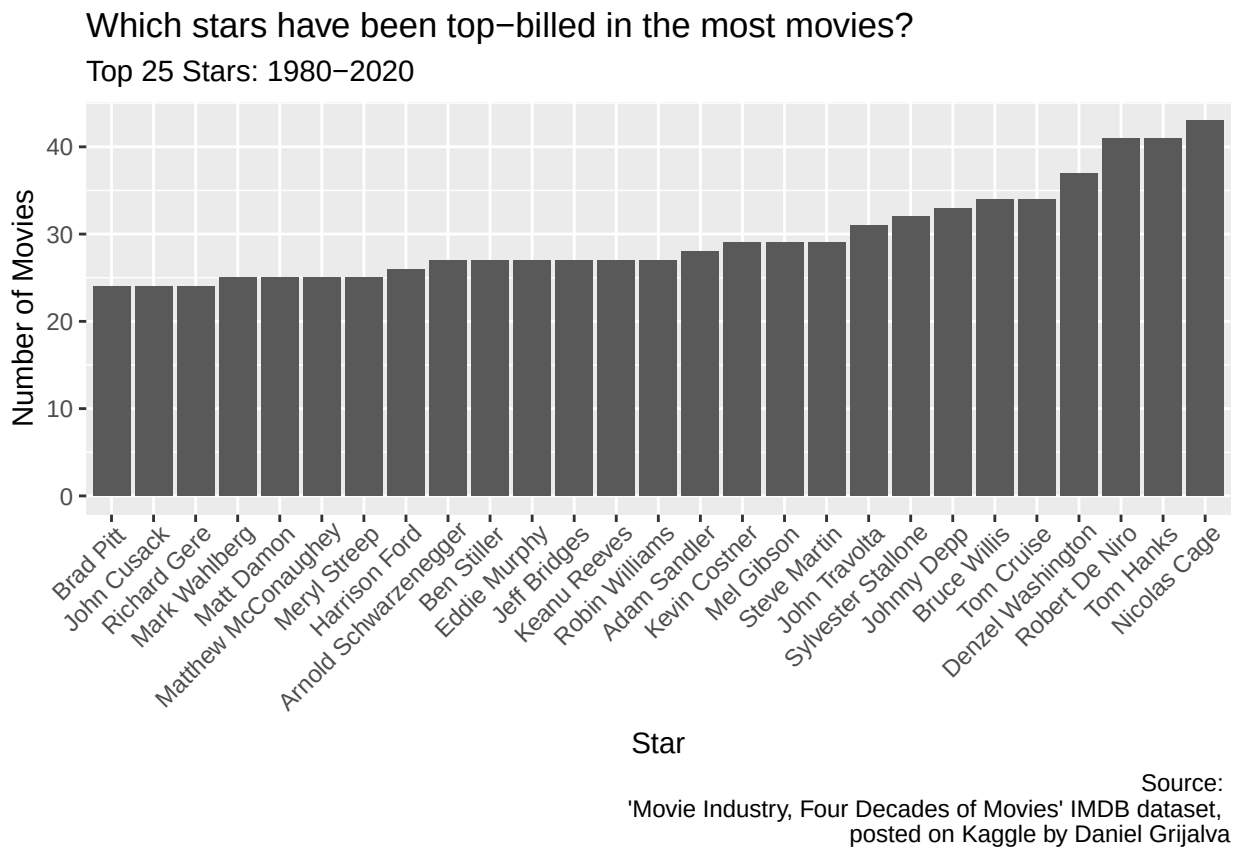
Total Range Analysis of Select Variables

Which stars have been top-billed in the most movies?

```
stars <- movies %>%
  count(star, sort = TRUE) %>%
  top_n(25)
```

Plot via column chart:

```
ggplot(stars) +
  geom_col(aes(x = reorder(star, n), y = n)) +
  labs(x = "Star", y = "Number of Movies",
       title = "Which stars have been top-billed in the most movies?",
       subtitle = "Top 25 Stars: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



Cage, De Niro, and Hanks are a clear Top 3.

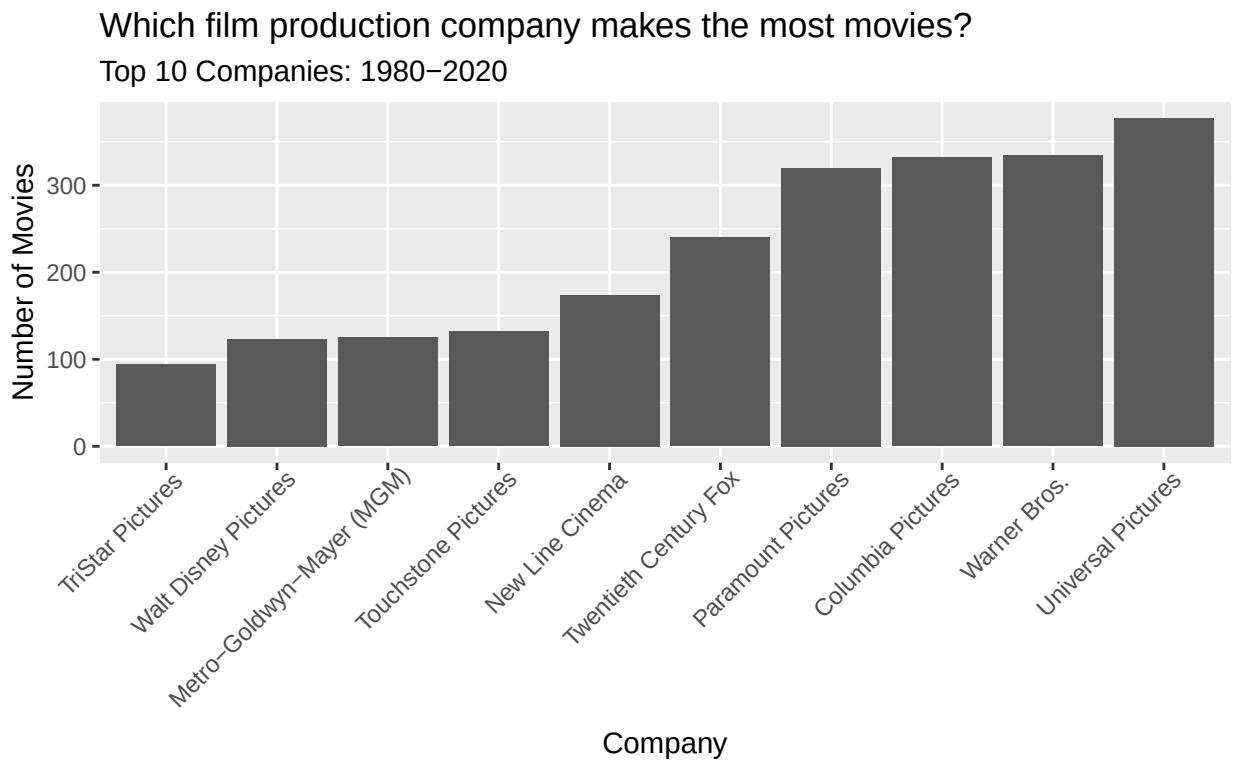
Note: I have to go all the way down to #24 before finding a female top-billed actor (Streep).

Which film production company made the most movies?

```
companies <- movies %>%
  count(company, sort = TRUE) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(companies) +
  geom_col(aes(x = reorder(company, n), y = n)) +
  labs(x = "Company", y = "Number of Movies",
       title = "Which film production company makes the most movies?",
       subtitle = "Top 10 Companies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



Universal, Warner Bros., Columbia, and Paramount are a clear Top 4 with over 300 films in the last 4 decades.

Universal appears multiple times on the Top 20 grossing list, and Warner Bros. appears once.

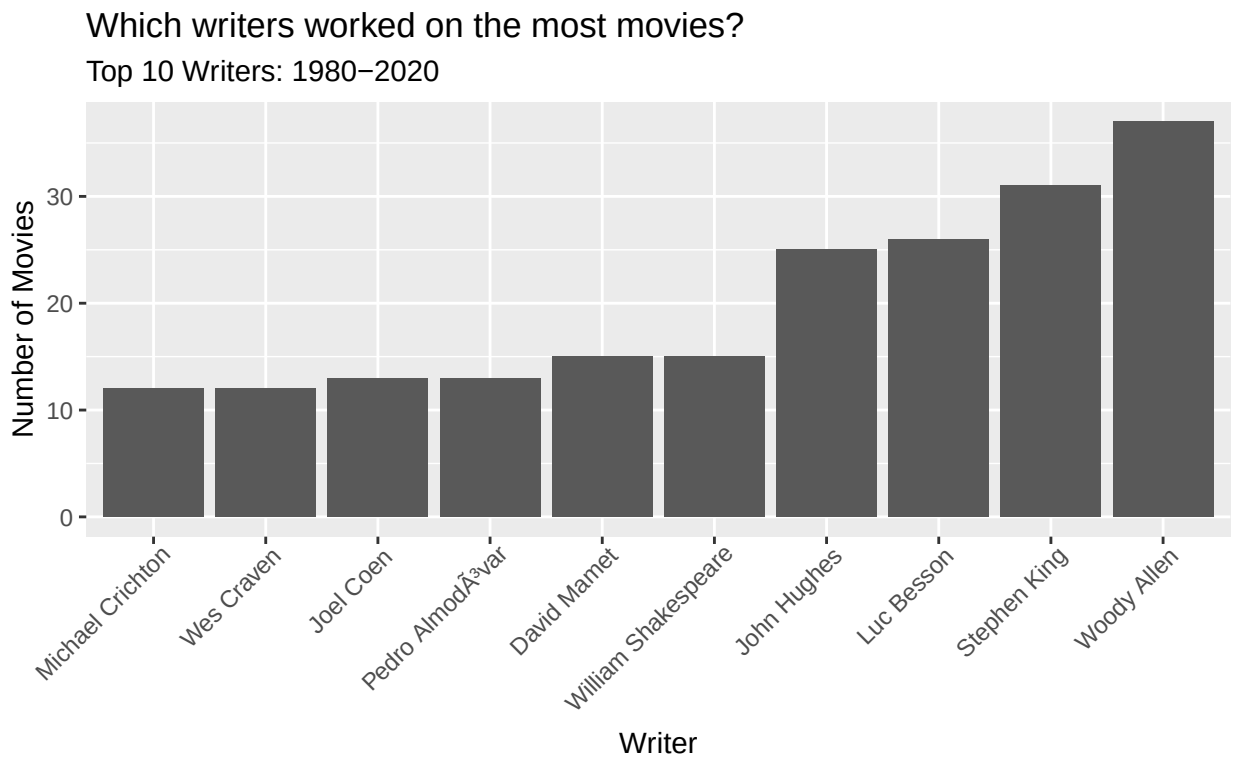
- Columbia and Paramount are not on the Top 20 grossing list.
- 20th Century and Walt Disney Pictures appear multiple times on the Top 20 grossing list, but made far less movies in the last 4 decades than the top 4 movie-making companies.

Which writers worked on the most movies?

```
writers <- movies %>%
  count(writer, sort = TRUE) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(writers) +
  geom_col(aes(x = reorder(writer, n), y = n)) +
  labs(x = "Writer", y = "Number of Movies",
       title = "Which writers worked on the most movies?",
       subtitle = "Top 10 Writers: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



At the top, Woody Allen and Stephen King have written over 30 movies, and Luc Besson and John Hughes have written over 20 movies.

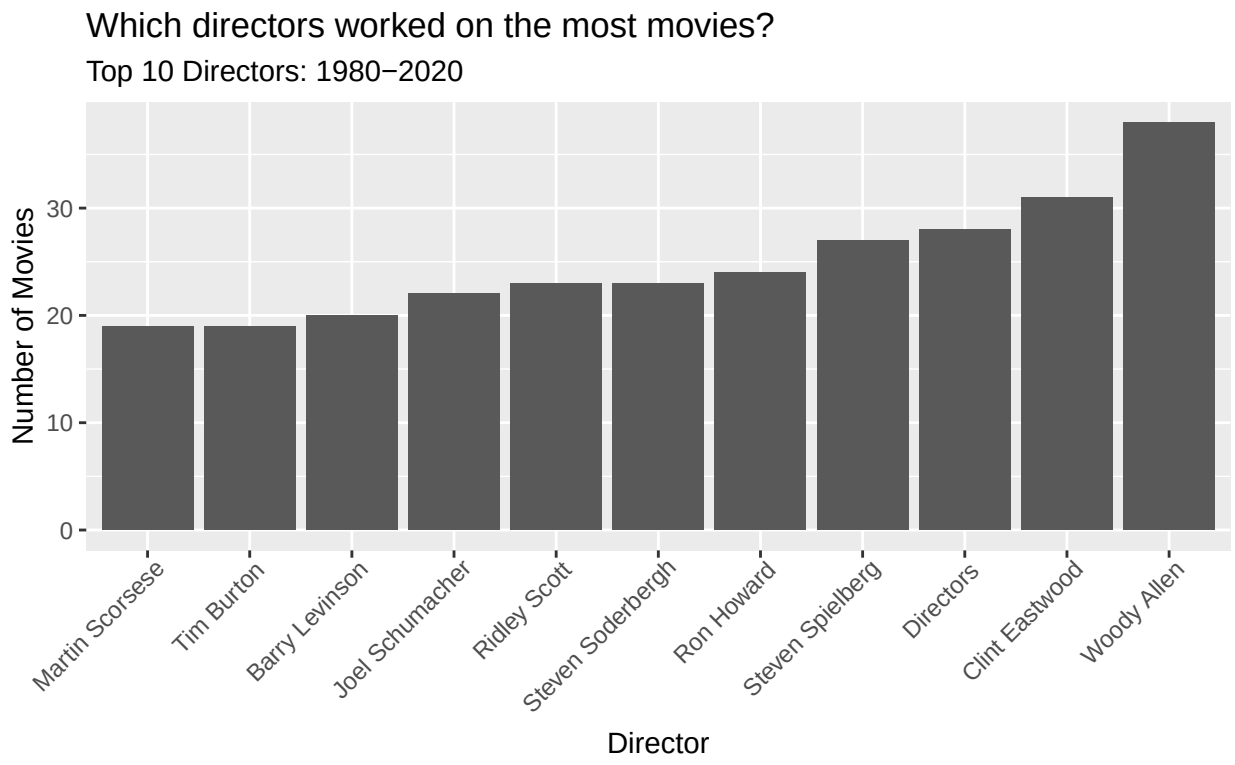
- None of the top 10 writers wrote the Top 20 grossing movies, nor the most profitable movies ($ or %).

Which directors worked on the most movies?

```
directors <- movies %>%
  count(director, sort = TRUE) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(directors) +
  geom_col(aes(x = reorder(director, n), y = n)) +
  labs(x = "Director", y = "Number of Movies",
       title = "Which directors worked on the most movies?",
       subtitle = "Top 10 Directors: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



At the top,Woody Allen and Clint Eastwood directed over 30 movies.

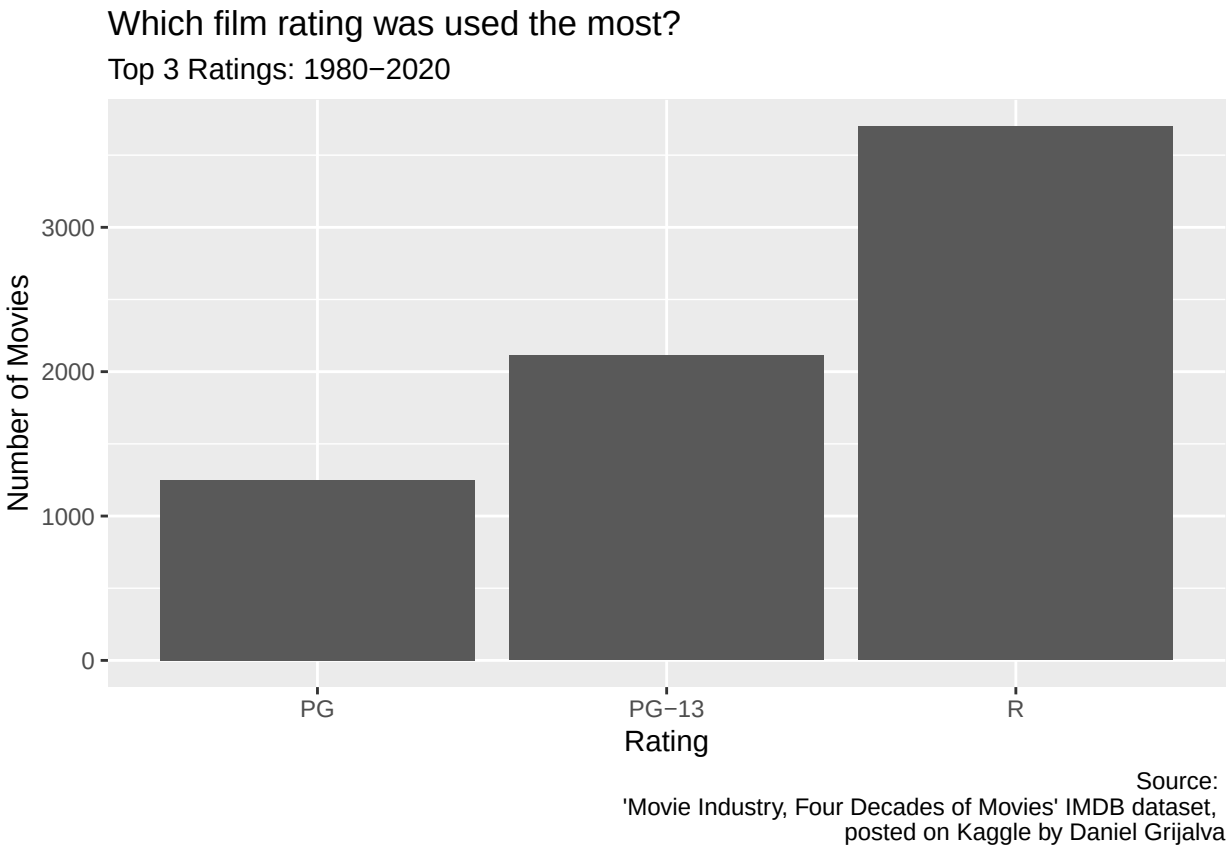
- None of the Top 10 directors wrote the Top 20 grossing movies, nor the most profitable movies ($ or %).

Which film rating was used the most?

```
ratings <- movies %>%
  count(rating, sort = TRUE) %>%
  top_n(3)
```

Plot via column chart:

```
ggplot(ratings) +
  geom_col(aes(x = reorder(rating, n), y = n)) +
  labs(x = "Rating", y = "Number of Movies",
       title = "Which film rating was used the most?",
       subtitle = "Top 3 Ratings: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



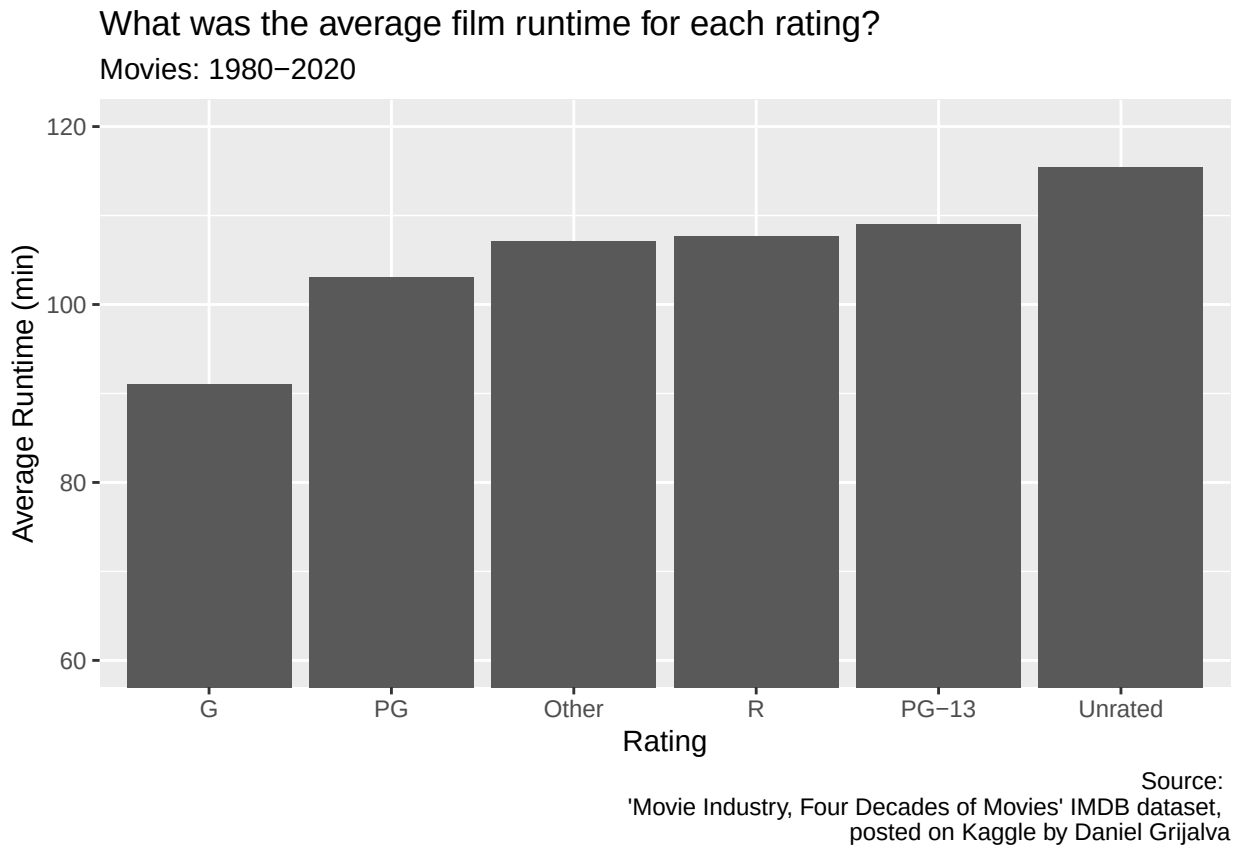
R rated films were made the most often (3698), followed by PG-13 (2112) and PG (1252).

What was the average film runtime for each film rating?

```
runtime <- movies %>%
  filter(!is.na(runtime), rating != "") %>%
  group_by(rating) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
```

Plot via column chart:

```
ggplot(runtime) +
  geom_col(aes(x = reorder(rating, AVGruntime), y = AVGruntime)) +
  labs(x = "Rating", y = "Average Runtime (min)",
       title = "What was the average film runtime for each rating?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  coord_cartesian(ylim = c(60, 120))
```



PG, TV-14, and G movies had the shortest runtime.

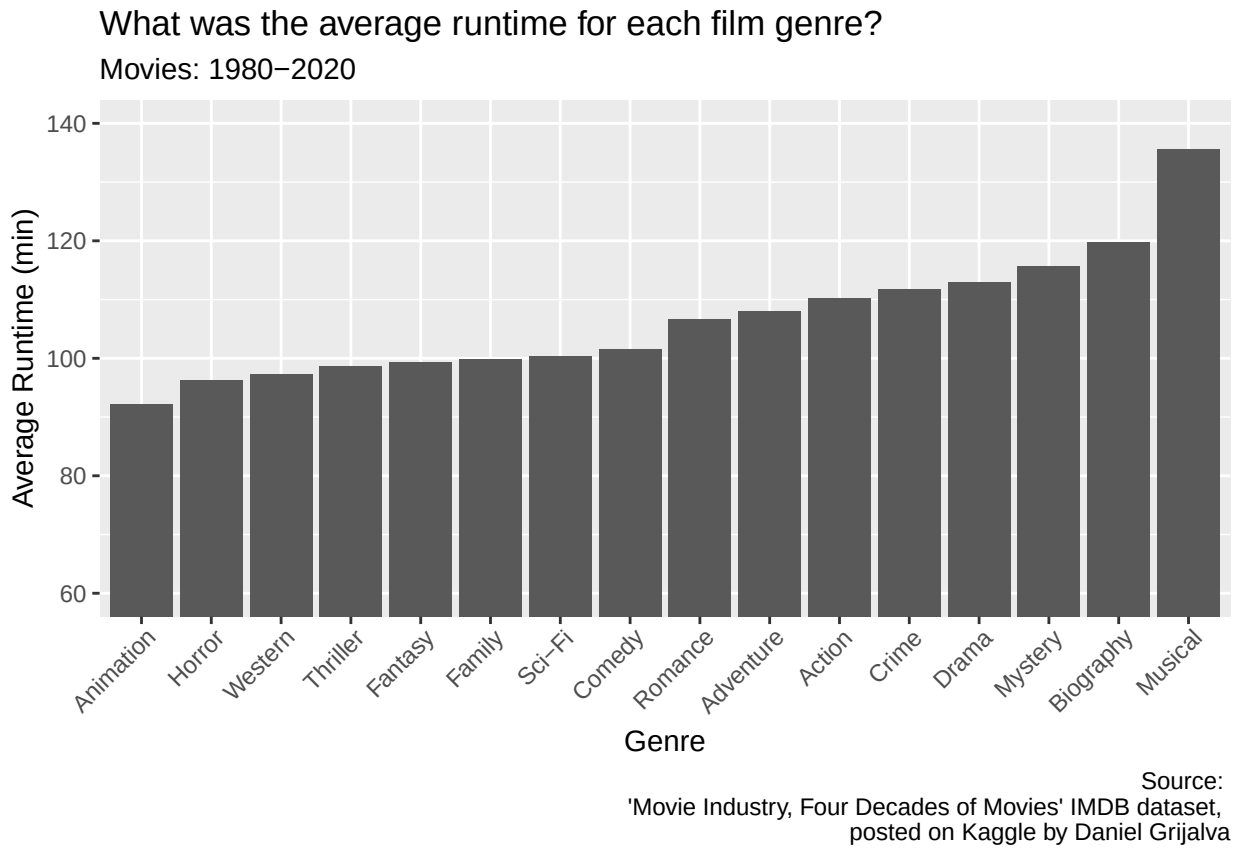
- PG and G make sense, since these are geared towards young children with shorter attention spans.

What was the average runtime for each film genre?

```
genre <- movies %>%
  filter(!is.na(runtime)) %>%
  group_by(genre) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
```

Plot via column chart:

```
ggplot(genre) +
  geom_col(aes(x = reorder(genre, AVGruntime), y = AVGruntime)) +
  labs(x = "Genre", y = "Average Runtime (min)",
       title = "What was the average runtime for each film genre?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  coord_cartesian(ylim = c(60, 140)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



Musicals were by far the longest (136 min). Animation was significantly lower at 92 min.

What companies made the most movies within each genre?

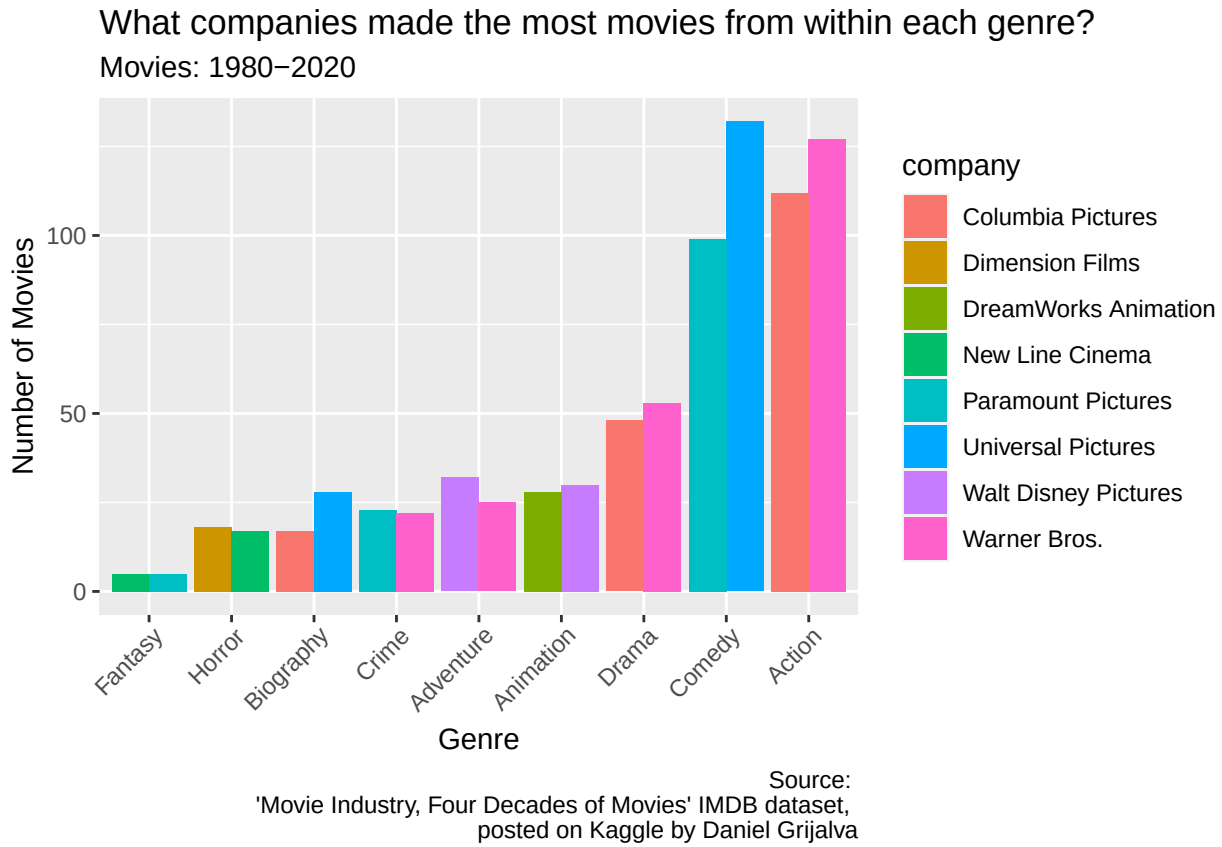
```
compGenre <- movies %>%
  group_by(genre, company) %>%
  count(genre, sort = TRUE)
compGenre <- compGenre %>%
  arrange(desc(n)) %>%
  group_by(genre) %>%
  slice(1:2)  # Top 2 highest values (number of movies made by company) by group (genre)
compGenre <- compGenre %>%  #some genres had < 10 movies, so I will cut them from the list
  filter(genre != "Family", genre != "Musical", genre != "Mystery", genre != "Romance",
         genre != "Sci-Fi", genre != "Thriller", genre != "Western")
print(compGenre)
```

```
## # A tibble: 18 x 3
## # Groups:   genre [9]
##    genre     company                 n
##    <chr>     <chr>               <int>
##  1 Action    Warner Bros.          127
##  2 Action    Columbia Pictures     112
##  3 Adventure Walt Disney Pictures   32
##  4 Adventure Warner Bros.           25
##  5 Animation Walt Disney Pictures   30
##  6 Animation DreamWorks Animation   28
##  7 Biography Universal Pictures     28
##  8 Biography Columbia Pictures      17
```

```
##  9 Comedy    Universal Pictures     132
## 10 Comedy    Paramount Pictures      99
## 11 Crime     Paramount Pictures      23
## 12 Crime     Warner Bros.            22
## 13 Drama     Warner Bros.            53
## 14 Drama     Columbia Pictures       48
## 15 Fantasy   New Line Cinema          5
## 16 Fantasy   Paramount Pictures       5
## 17 Horror    Dimension Films         18
## 18 Horror    New Line Cinema         17
```

Plot via clustered column chart:

```
ggplot(compGenre) +
  geom_col(aes(x = reorder(genre, n), y = n, fill = company), position = "dodge") +
  labs(x = "Genre", y = "Number of Movies",
       title = "What companies made the most movies from within each genre?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



Warner Bros. made the most Action movies (127).

Universal Pictures made the most Comedy movies (132).

Top Grossing Analysis of Select Variables

Top 20 Highest Grossing Movies:

```
top20gross <- movies %>%
  select(c(name, rating, runtime, genre, profitM, profit_percent, grossM)) %>%
  arrange(desc(grossM)) %>%
  top_n(20)
top20gross
```

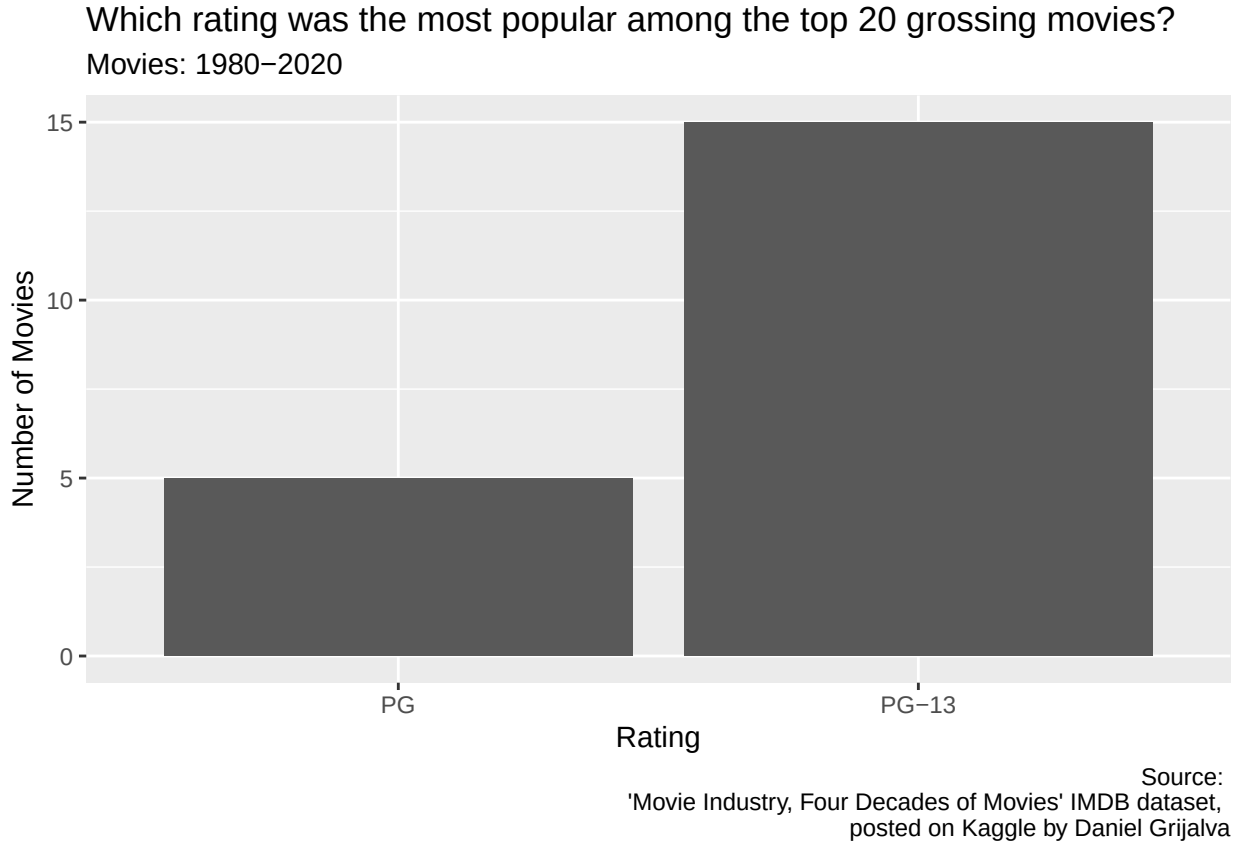
```
##                                         name rating runtime     genre
## 1                                     Avatar  PG-13     162    Action
## 2                          Avengers: Endgame  PG-13     181    Action
## 3                                    Titanic  PG-13     194     Drama
## 4    Star Wars: Episode VII - The Force Awakens  PG-13     138    Action
## 5                     Avengers: Infinity War  PG-13     149    Action
## 6                              The Lion King     PG     118 Animation
## 7                            Jurassic World  PG-13     124    Action
## 8                              The Avengers  PG-13     143    Action
## 9                                 Furious 7  PG-13     137    Action
## 10                                Frozen II     PG     103 Animation
## 11                    Avengers: Age of Ultron  PG-13     141    Action
## 12                             Black Panther  PG-13     134    Action
## 13 Harry Potter and the Deathly Hallows: Part 2  PG-13     130 Adventure
## 14      Star Wars: Episode VIII - The Last Jedi  PG-13     152    Action
## 15               Jurassic World: Fallen Kingdom  PG-13     128    Action
## 16                                    Frozen     PG     102 Animation
## 17                       Beauty and the Beast     PG     129    Family
## 18                             Incredibles 2     PG     118 Animation
## 19                    The Fate of the Furious  PG-13     136    Action
## 20                                Iron Man 3  PG-13     130    Action
##      profitM profit_percent   grossM
## 1  2610.2462      1101.3697 2847.246
## 2  2441.5013       685.8150 2797.501
## 3  2001.6473      1000.8236 2201.647
## 4  1824.5217       744.7027 2069.522
## 5  1727.3598       538.1183 2048.360
## 6  1410.7276       542.5875 1670.728
## 7  1520.5164      1013.6776 1670.516
## 8  1298.8155       590.3707 1518.816
## 9  1325.3414       697.5481 1515.341
## 10 1300.0269       866.6846 1450.027
## 11 1152.8095       461.1238 1402.810
## 12 1147.5980       573.7990 1347.598
## 13 1217.3217       973.8573 1342.322
## 14 1015.6988       320.4097 1332.699
## 15 1140.4663       670.8625 1310.466
## 16 1131.5081       754.3387 1281.508
## 17 1104.4345       690.2716 1264.435
## 18 1044.6395       522.3198 1244.640
## 19  986.0051       394.4020 1236.005
## 20 1014.8113       507.4056 1214.811
```

Which rating was the most popular among top grossing movies?

```
ratingtop20gross <- top20gross %>%
  count(rating, sort = TRUE)
```

Plot via column chart:

```
ggplot(ratingtop20gross, aes(rating, n)) +
  geom_col() +
  labs(x = "Rating", y = "Number of Movies",
       title = "Which rating was the most popular among the top 20 grossing movies?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



PG-13 movies dominated the Top 20 Highest Grossing Movie list with 15.

PG movies have 5, and R movies have 0.

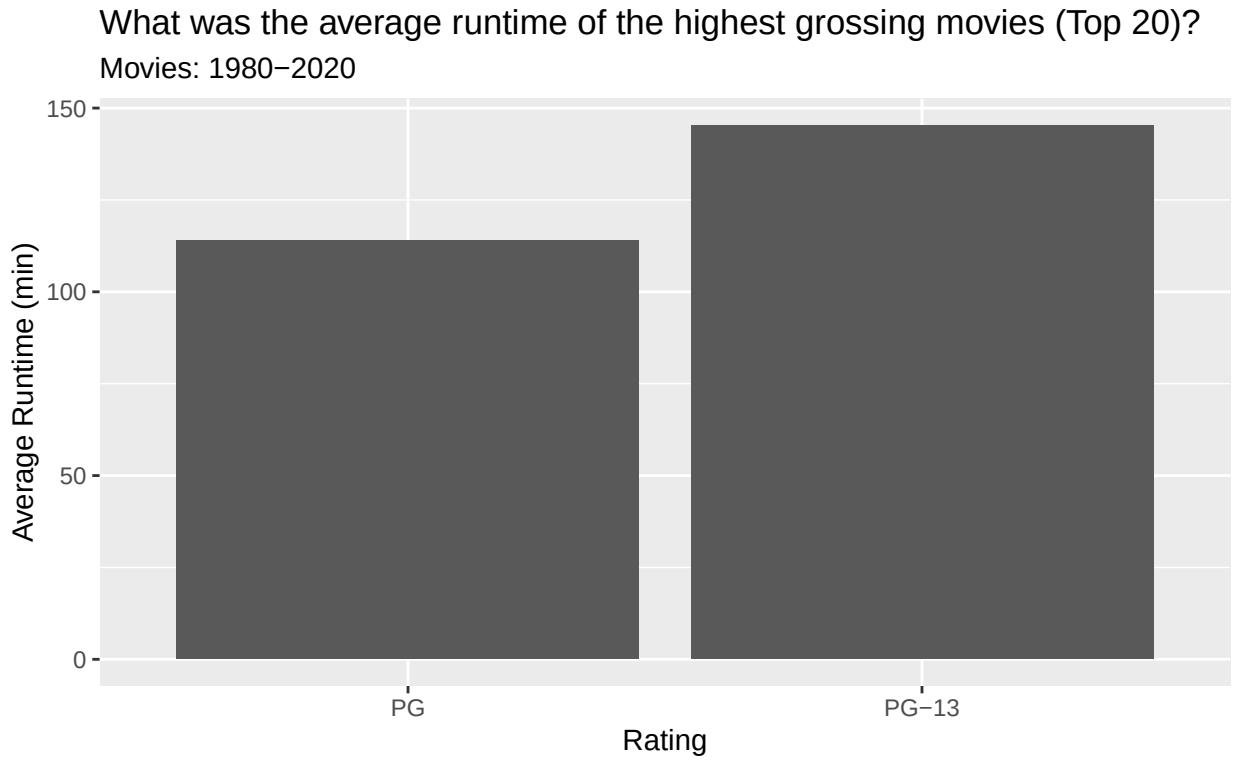
What was the average runtime of the highest grossing movies (Top 20)?

```
runtimetop20gross <- top20gross %>%
  group_by(rating) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
```

Plot via column chart:

```
ggplot(runtimetop20gross, aes(rating, AVGruntime)) +
  geom_col() +
```

```
  labs(x = "Rating", y = "Average Runtime (min)",
       title = "What was the average runtime of the highest grossing movies (Top 20)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



For PG-13, the avg runtime of the top 20 grossing movies was 145 min, which was 36 min longer than the avg of all PG-13 movies.

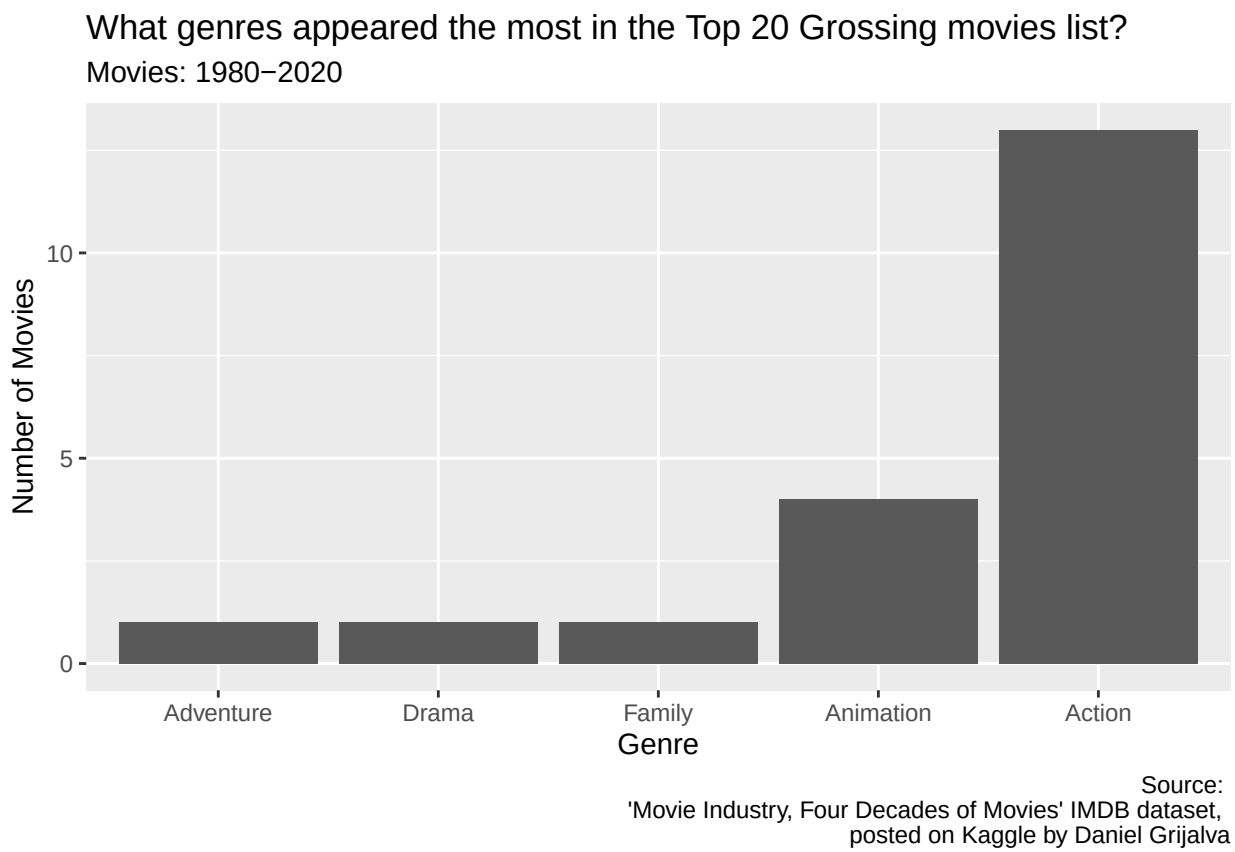
For PG, the ave runtime of the top 20 grossing movies was 114 min, which was 9 min longer than the avg of all PG movies.

What genres appeared the most in the Top 20 Grossing movies list?

```
genretop20gross <- top20gross %>%
  group_by(genre) %>%
  count(genre, sort = TRUE)
```

Plot via column chart:

```
ggplot(genretop20gross, aes(x = reorder(genre, n), y = n)) +
  geom_col() +
  labs(x = "Genre", y = "Number of Movies",
       title = "What genres appeared the most in the Top 20 Grossing movies list?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



Action movies accounted for 13 out of the top 20 grossing films.

Top Profitable Analysis of Select Variables

Top 20 Movies with Highest Profit Percentage:

```
top20profperc <- movies %>%
  select(c(name, rating, runtime, genre, profitM, profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(20)
top20profperc
```

```
##                         name rating runtime     genre    profitM
## 1         Paranormal Activity      R      86    Horror 193.340800
## 2     The Blair Witch Project      R      81    Horror 248.579099
## 3                   The Gallows      R      81    Horror  42.864410
## 4                   El Mariachi      R      81    Action   2.033920
## 5                          Once      R      86     Drama  20.786722
## 6                        Clerks      R      92    Comedy   3.124130
## 7            Napoleon Dynamite     PG      96    Comedy  45.738887
## 8       In the Company of Men      R      97    Comedy   2.779473
## 9                   Keeping Mum      R      99    Comedy  18.417834
## 10                  Open Water      R      79 Adventure  54.183487
## 11            The Devil Inside      R      83    Horror 100.758490
## 12              The Quiet Ones  PG-13      98    Horror  17.635162
## 13                         Saw      R     103    Horror 102.711669
```

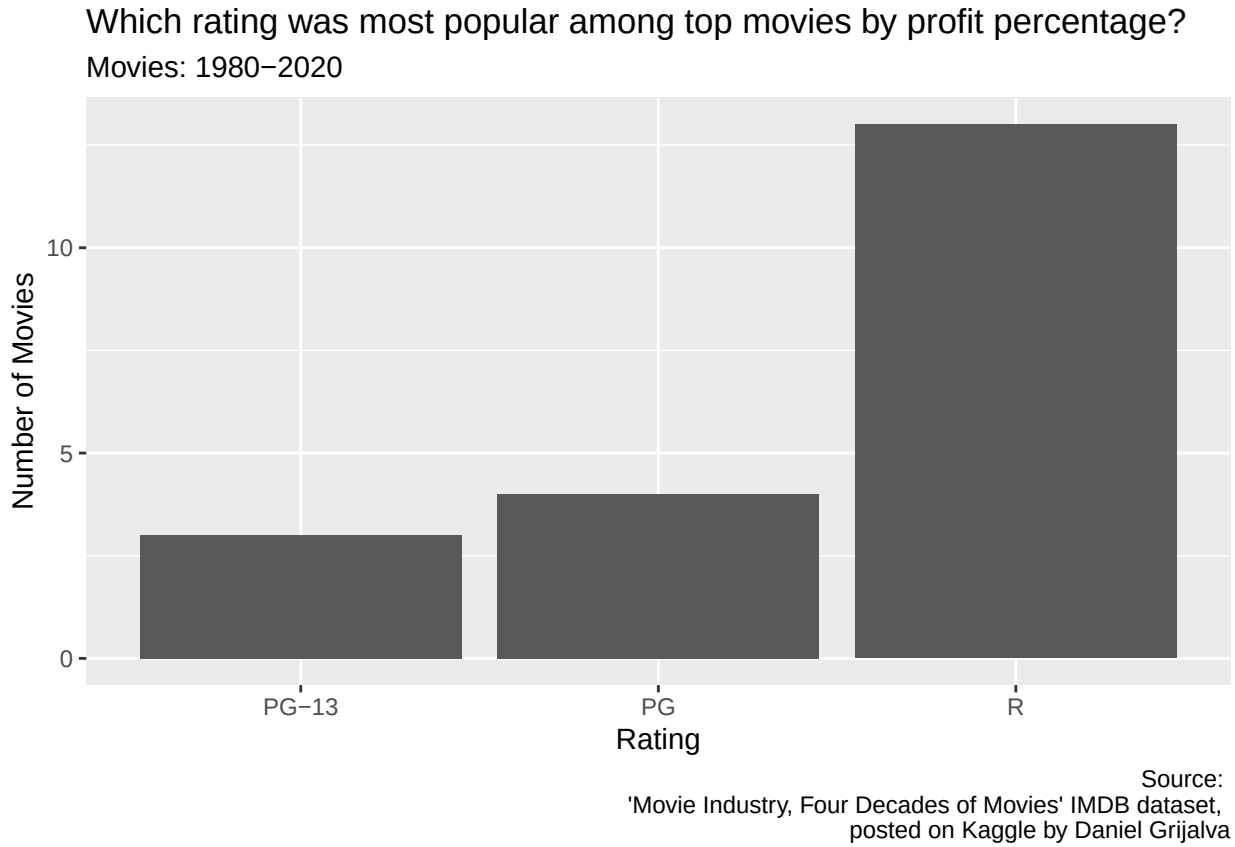
```
## 14                  Searching  PG-13     102     Drama  74.582037
## 15                     Primer  PG-13      77     Drama   0.538436
## 16 E.T. the Extra-Terrestrial     PG     115    Family 782.410554
## 17   My Big Fat Greek Wedding     PG      95    Comedy 363.744044
## 18              The Full Monty      R      91    Comedy 254.438649
## 19            Friday the 13th       R      95    Horror  39.204601
## 20                  Fireproof     PG     122     Drama  32.973297
##    profit_percent
## 1     1288938.667
## 2      414298.498
## 3       42864.410
## 4       29056.000
## 5       13857.815
## 6       11570.852
## 7       11434.722
## 8       11117.892
## 9       10898.127
## 10      10836.697
## 11      10075.849
## 12       8817.581
## 13       8559.306
## 14       8475.231
## 15       7691.943
## 16       7451.529
## 17       7274.881
## 18       7269.676
## 19       7128.109
## 20       6594.659
```

Which rating was most popular among top movies by profit percentage?

```
ratingtop20profperc <- top20profperc %>%
  count(rating, sort = TRUE)
```

Plot via column chart:

```
ggplot(ratingtop20profperc, aes(x = reorder(rating, n), y = n)) +
  geom_col() +
  labs(x = "Rating", y = "Number of Movies",
       title = "Which rating was most popular among top movies by profit percentage?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



R movies dominated the top profit (%) movie list with 13.

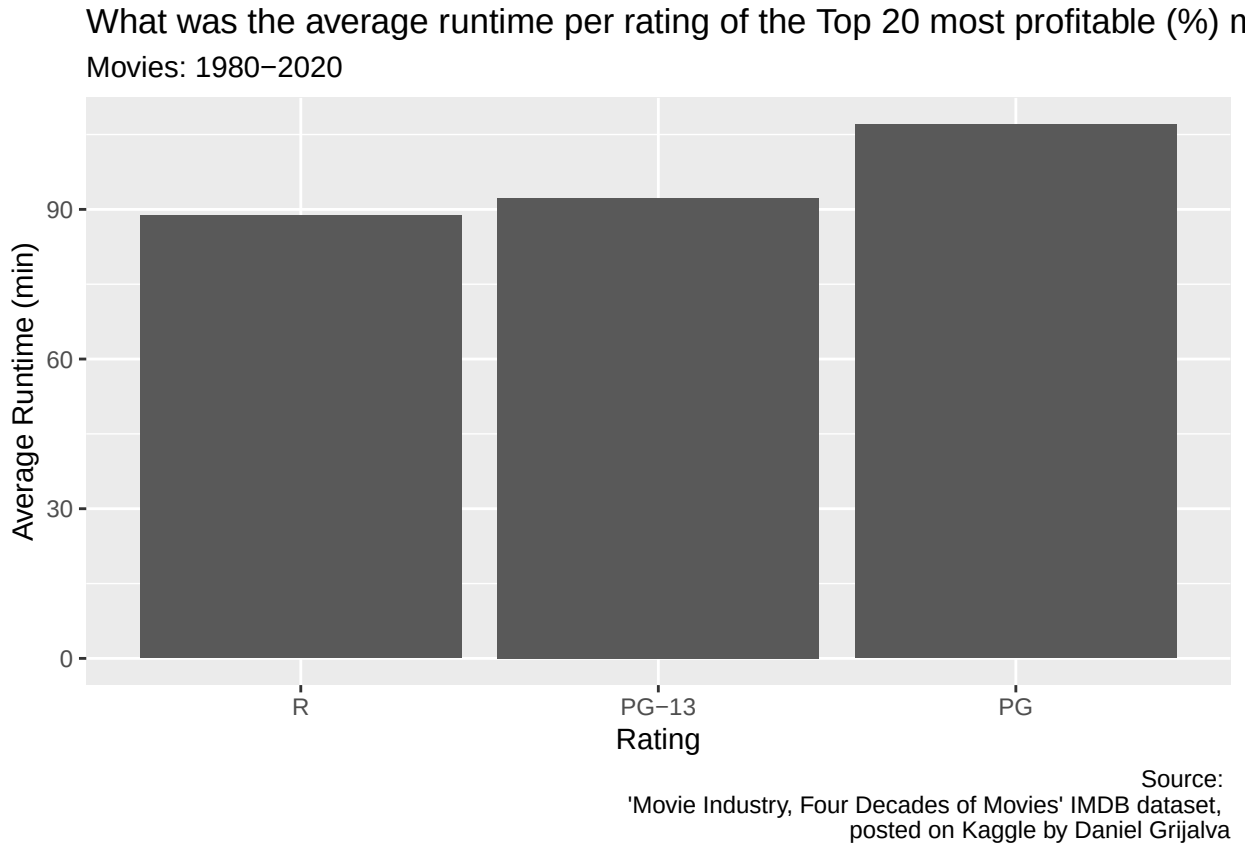
PG movies had 4, and PG-13 movies had 3.

What was the average runtime per rating of the Top 20 most profitable (%) movies?

```
runtimetop20profperc <- top20profperc %>%
  group_by(rating) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
```

Plot via column chart:

```
ggplot(runtimetop20profperc, aes(x = reorder(rating, AVGruntime), y = AVGruntime)) +
  geom_col() +
  labs(x = "Rating", y = "Average Runtime (min)",
       title =
         "What was the average runtime per rating of the Top 20 most profitable (%) movies?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



For PG, the avg runtime of the top 20 most profitable (%) movies was 107 min, which was 4 min longer than the avg of all PG movies.

For PG-13, the avg runtime of the top 20 most profitable (%) movies was 92 min, which was 17 min less than the avg of all PG-13 movies.

For R, the avg runtime of the top 20 most profitable (%) movies was 89 min, which was 19 min less than the avg of all R movies.

Insights: top grossing movies were significantly longer compared to all movies with the same ratings.

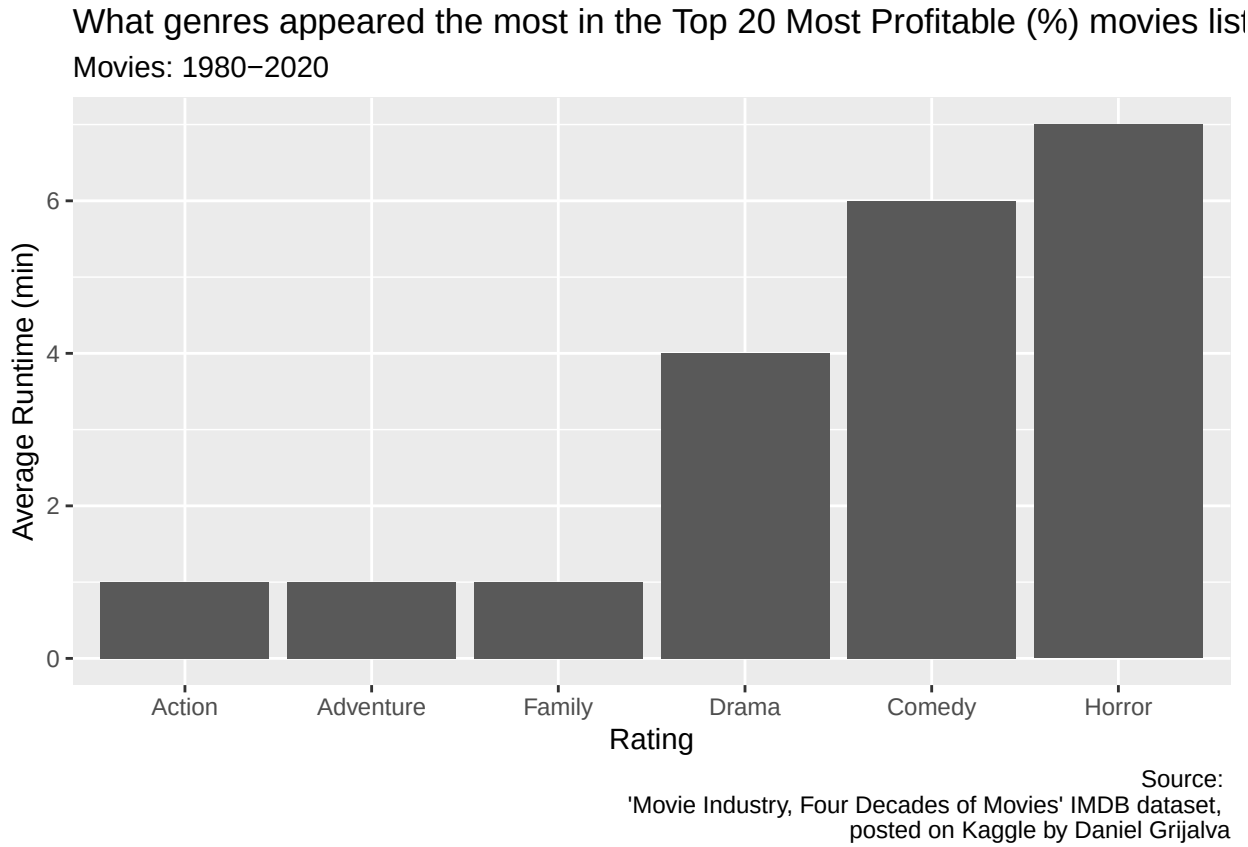
- Top profitable (%) movies were generally shorter than the average, perhaps due to smaller budgets.

What genres appeared the most in the Top 20 Most Profitable (%) movies list?

```
genretop20profperc<- top20profperc %>%
  group_by(genre) %>%
  count(genre, sort = TRUE)
```

Plot via column chart:

```
ggplot(genretop20profperc, aes(x = reorder(genre, n), y = n)) +
  geom_col() +
  labs(x = "Rating", y = "Average Runtime (min)",
       title =
         "What genres appeared the most in the Top 20 Most Profitable (%) movies list?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



Horror had the most (7), followed by Comedy (6) and Drama (4).

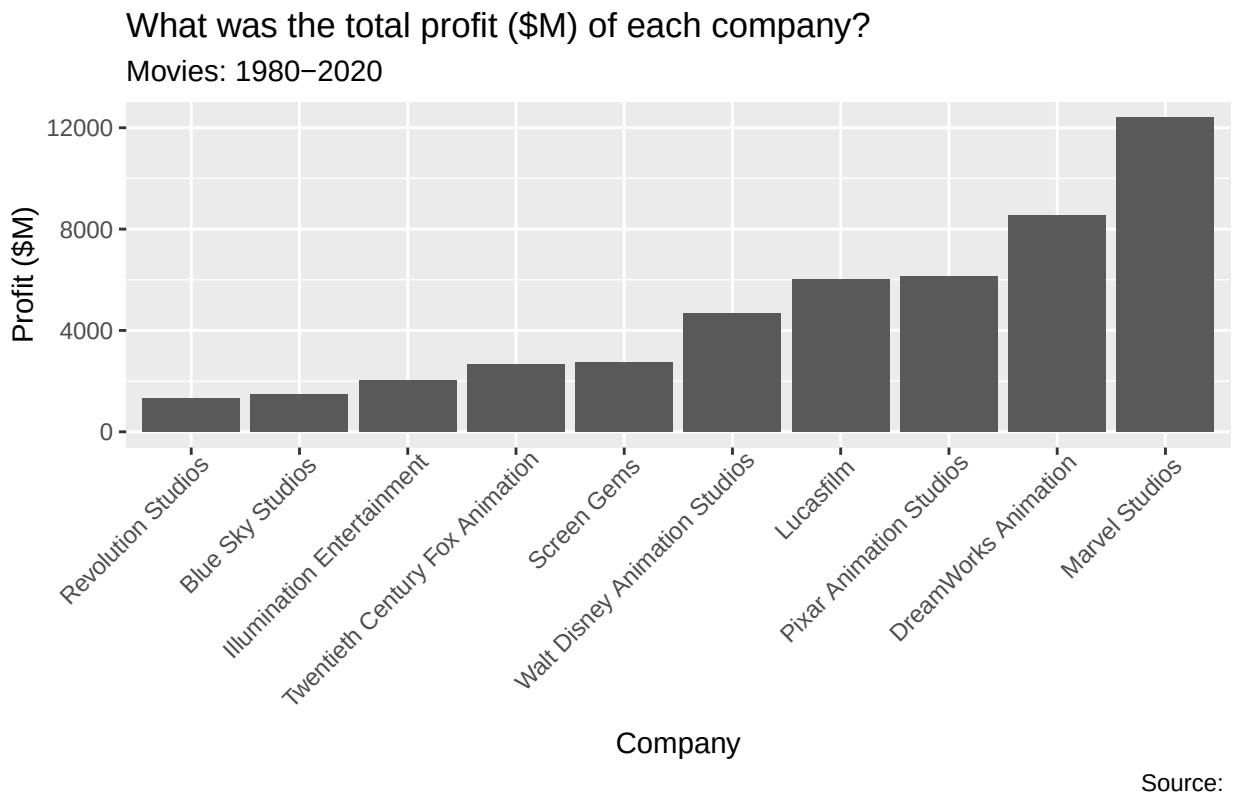
- Only 1 Action film, which contrasts with the top grossing genres.

What was the total profit ($M) of each company?

```
top10profitcomp <- movies %>%
  group_by(company) %>%
  summarise(profitM = sum(profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(top10profitcomp, aes(x = reorder(company, profitM), y = profitM)) +
  geom_col() +
  labs(x = "Company", y = "Profit ($M)",
       title = "What was the total profit ($M) of each company?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



Marvel had the most profit by far ($12B), followed by Dreamworks ($8B), Pixar ($6B), and Lucasfilm ($6B).

Avg profit per movie of each company (with at least 5 movies)

Use companies that have made at least 5 movies:

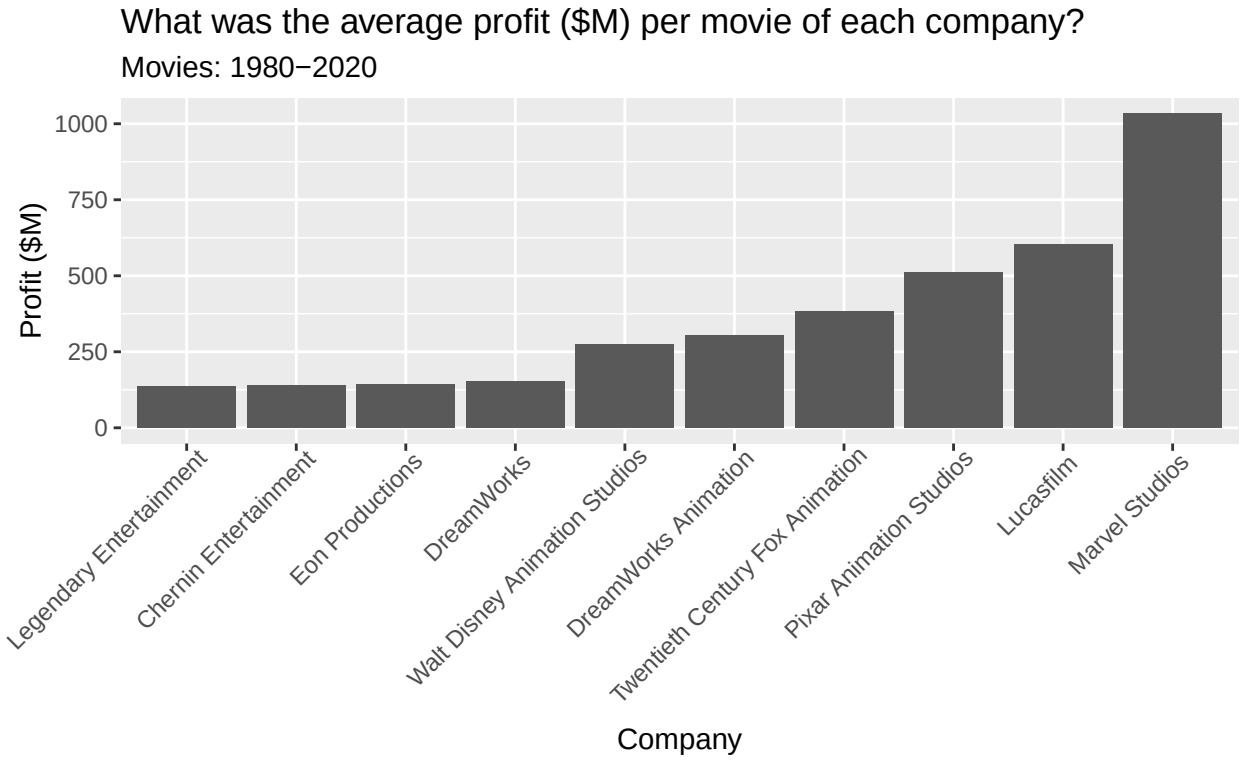
```
compN <- movies %>%
  group_by(company) %>% filter(n() >= 5) %>% ungroup()
```

What was the average profit ($M) per movie of each company?

```
profcompN <- compN %>%
  group_by(company) %>%
  summarise(profitM = mean(profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(profcompN, aes(x = reorder(company, profitM), y = profitM)) +
  geom_col() +
  labs(x = "Company", y = "Profit ($M)",
       title = "What was the average profit ($M) per movie of each company?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



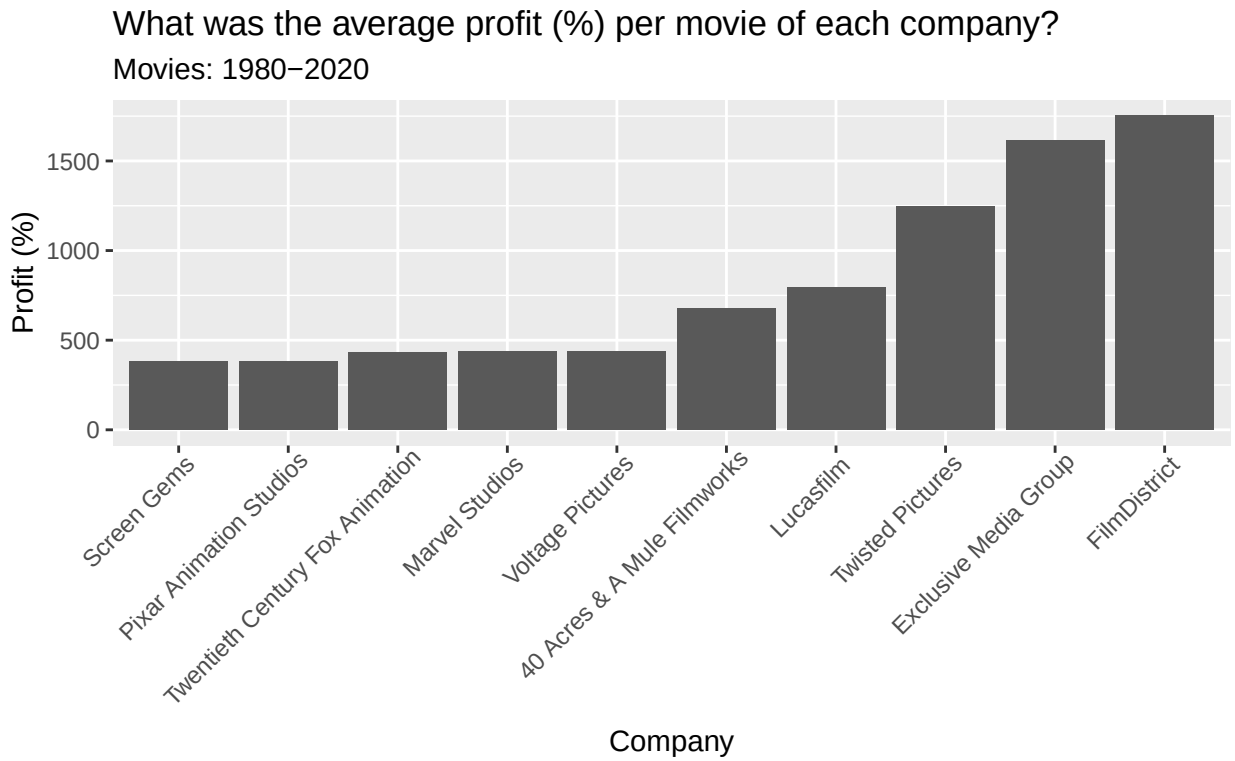
Marvel Studios made about $1B profit per movie.

What was the average profit (%) per movie of each company?

```
perccompN <- compN %>%
  group_by(company) %>%
  summarise(profit_percent = mean(profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(10)
```

Plot via column chart:

```
ggplot(perccompN, aes(x = reorder(company, profit_percent), y = profit_percent)) +
  geom_col() +
  labs(x = "Company", y = "Profit (%)",
       title = "What was the average profit (%) per movie of each company?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.95))
```



FilmDistrict, Exclusive Media Group, and Twisted Pictures all averaged over 1000% profit from their movies.

Decade Analysis of Select Variables

Note: For Decade Analysis, I did not include 2020 films.

How many movies did companies make each decade (80s, 90s, 00s, 10s)?

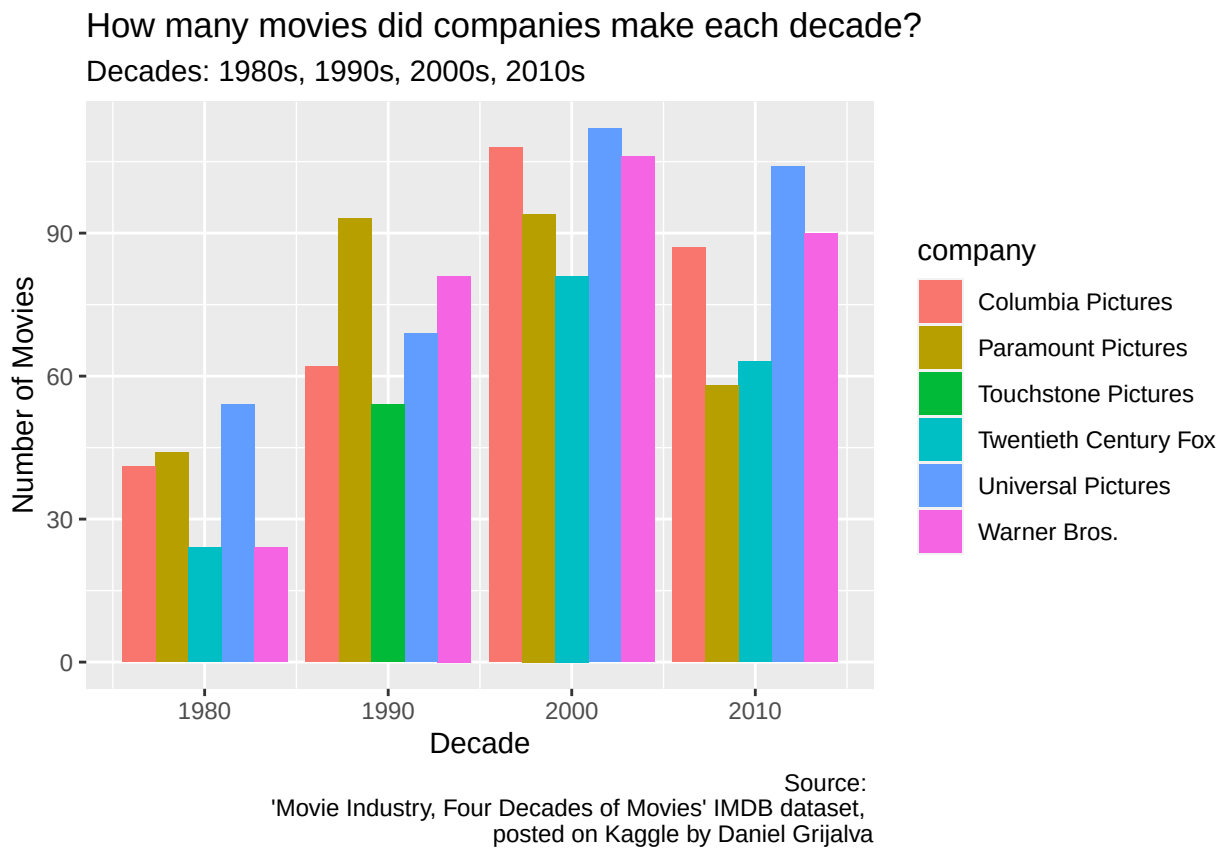
```
compDecade <- movies %>%
  filter(!is.na(decade), decade != 2020) %>% # eliminate NA decades and 2020 films
  group_by(decade, company) %>%
  count(decade, sort = TRUE)
compDecade <- compDecade %>%
  arrange(desc(n)) %>%
  group_by(decade) %>%
  slice(1:5)  # Top 5 highest values (number of movies made by company) by group (decade)
compDecade
```

```
## # A tibble: 20 x 3
## # Groups:   decade [4]
##    decade company                 n
##     <dbl> <chr>                 <int>
##  1   1980 Universal Pictures       54
##  2   1980 Paramount Pictures       44
##  3   1980 Columbia Pictures        41
##  4   1980 Twentieth Century Fox    24
##  5   1980 Warner Bros.             24
```

```
##  6   1990 Paramount Pictures       93
##  7   1990 Warner Bros.             81
##  8   1990 Universal Pictures       69
##  9   1990 Columbia Pictures        62
## 10   1990 Touchstone Pictures      54
## 11   2000 Universal Pictures      112
## 12   2000 Columbia Pictures       108
## 13   2000 Warner Bros.            106
## 14   2000 Paramount Pictures       94
## 15   2000 Twentieth Century Fox    81
## 16   2010 Universal Pictures      104
## 17   2010 Warner Bros.            90
## 18   2010 Columbia Pictures       87
## 19   2010 Twentieth Century Fox    63
## 20   2010 Paramount Pictures       58
```

Plot via clustered column chart:

```
ggplot(compDecade) +
  geom_col(aes(x = decade, y = n, fill = company), position = "dodge") +
  labs(x = "Decade", y = "Number of Movies",
       title = "How many movies did companies make each decade?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```

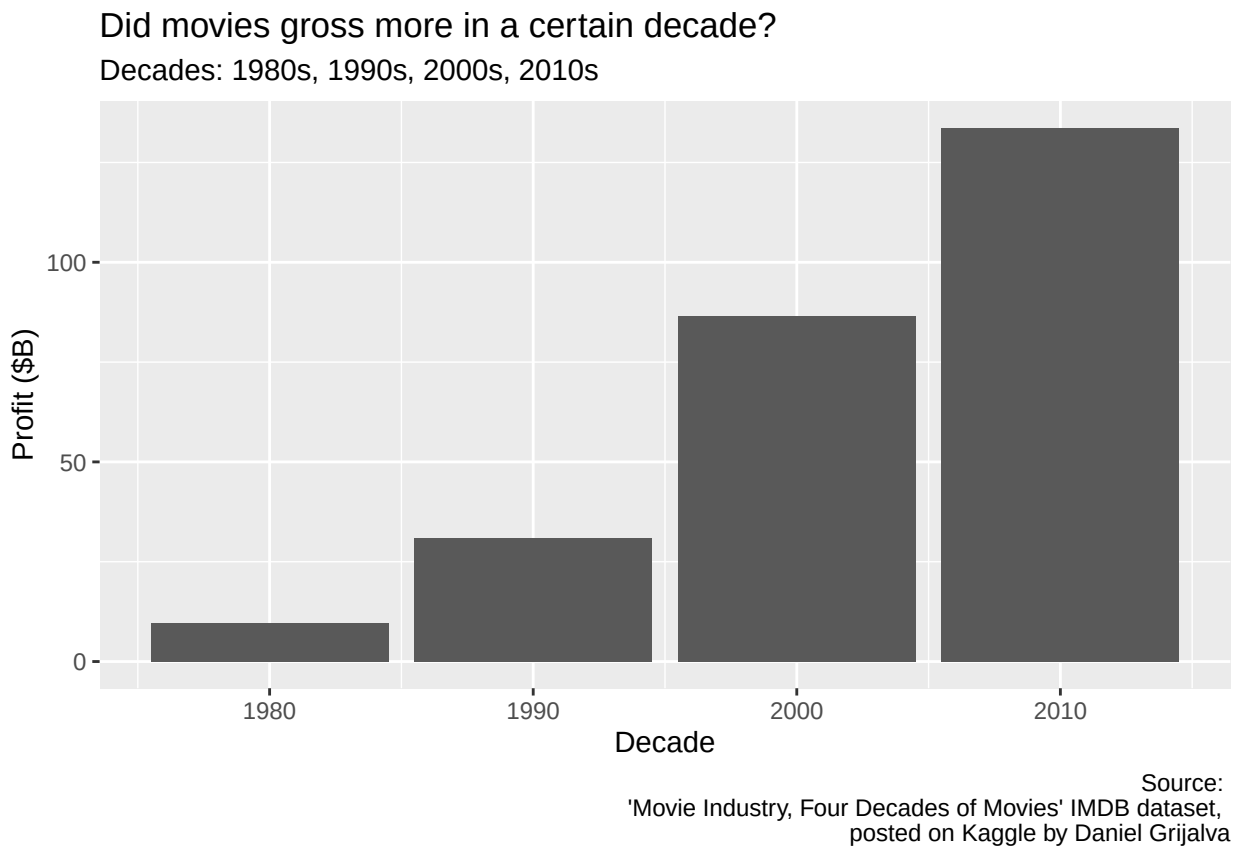


Did movies gross more in a certain decade?

```
profdecade <- movies %>%
  filter(!is.na(profitM), decade != 2020) %>%
  group_by(decade) %>%
  summarise(profitB = sum(profitM) / 1000) %>%
  arrange(desc(profitB))
```

Plot via column chart:

```
ggplot(profdecade, aes(x = decade, y = profitB)) +
  geom_col() +
  labs(x = "Decade", y = "Profit ($B)",
       title = "Did movies gross more in a certain decade?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



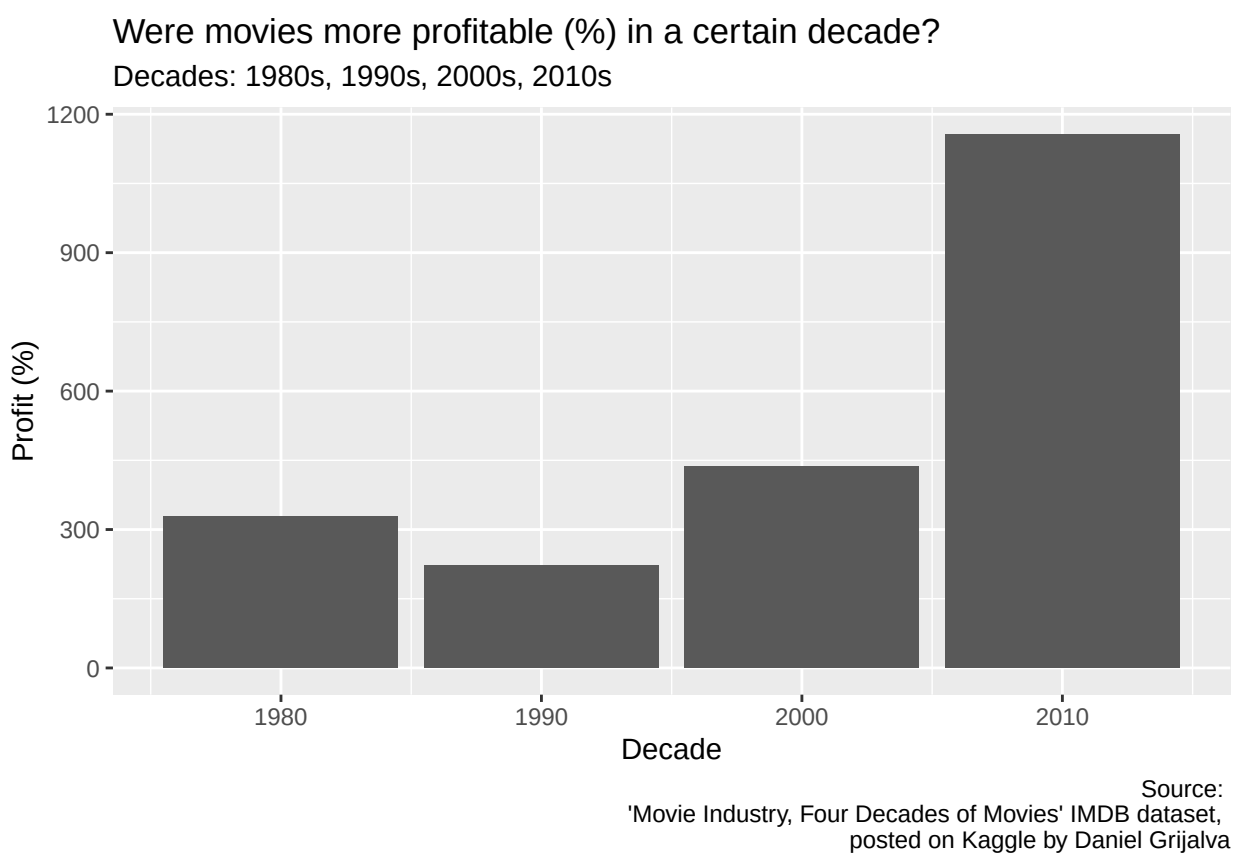
There was a cumulative rise in gross each decade. The cumulative rise in movie ticket prices is one possible factor.

Were movies more profitable (%) in a certain decade?

```
percdecade <- movies %>%
  filter(!is.na(profit_percent), decade != 2020) %>%
  group_by(decade) %>%
  summarise(profit_percent = mean(profit_percent)) %>%
  arrange(desc(profit_percent))
```

Plot via column chart:

```
ggplot(percdecade, aes(x = decade, y = profit_percent)) +
  geom_col() +
  labs(x = "Decade", y = "Profit (%)",
       title = "Were movies more profitable (%) in a certain decade?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



Movies were significantly more profitable (%) in the 2010s (1157%), followed by 2000s (437%), 1980s (329%), and 1990s (222%).

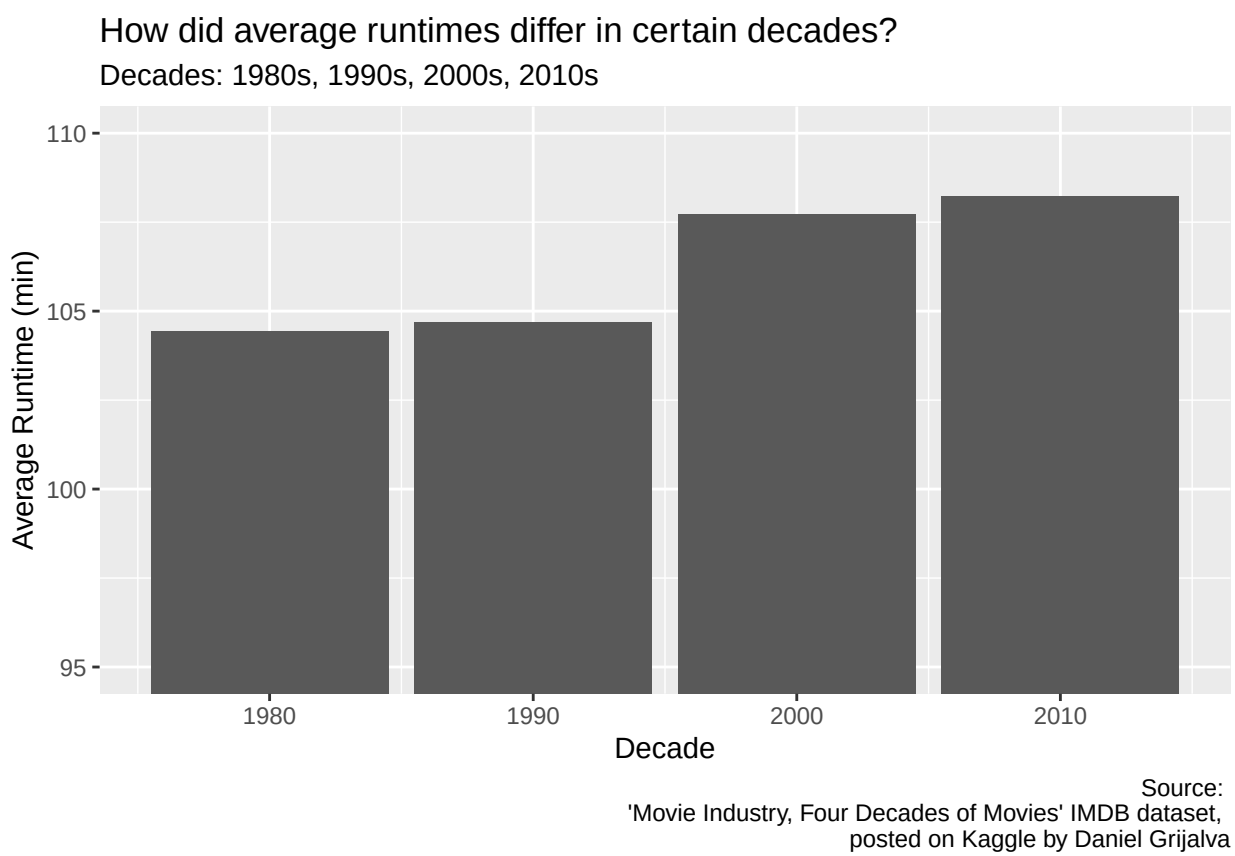
How did average runtimes differ in certain decades?

```
runtimedecade <- movies %>%
  filter(!is.na(runtime), decade != 2020) %>%
  group_by(decade) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
```

Plot via column chart:

```
ggplot(runtimedecade, aes(x = decade, y = AVGruntime)) +
  geom_col() +
  labs(x = "Decade", y = "Average Runtime (min)",
       title = "How did average runtimes differ in certain decades?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
```

```
      caption = "Source:
      'Movie Industry, Four Decades of Movies' IMDB dataset,
      posted on Kaggle by Daniel Grijalva") +
  coord_cartesian(ylim = c(95, 110))
```



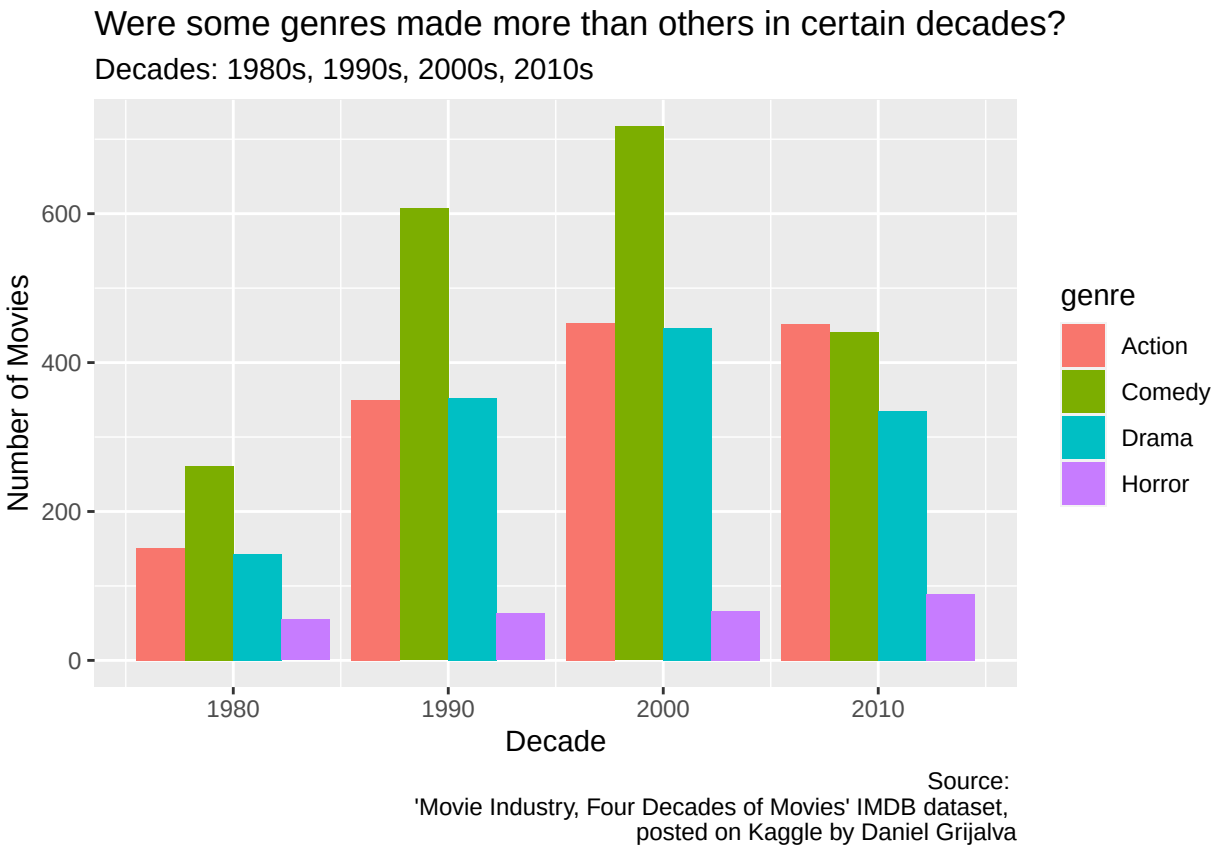
On average, movies get longer every decade.

Were some genres made more than others in certain decades?

```
genredecade <- movies %>%
  filter(!is.na(genre), decade != 2020) %>%
  filter(genre == "Comedy" | genre == "Action" | genre == "Drama" | genre == "Horror") %>%
  group_by(decade) %>%
  count(genre, sort = TRUE) %>%
  arrange(genre)
```

Plot via clustered column chart:

```
ggplot(genredecade) +
  geom_col(aes(x = decade, y = n, fill = genre), position = "dodge") +
  labs(x = "Decade", y = "Number of Movies",
       title = "Were some genres made more than others in certain decades?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



More action movies were made in 00s-10s than in 80s-90s.

More comedy movies were made in 90s-00s than in 80s/10s.

More drama movies were made in 90s-00s than in 80s/10s.

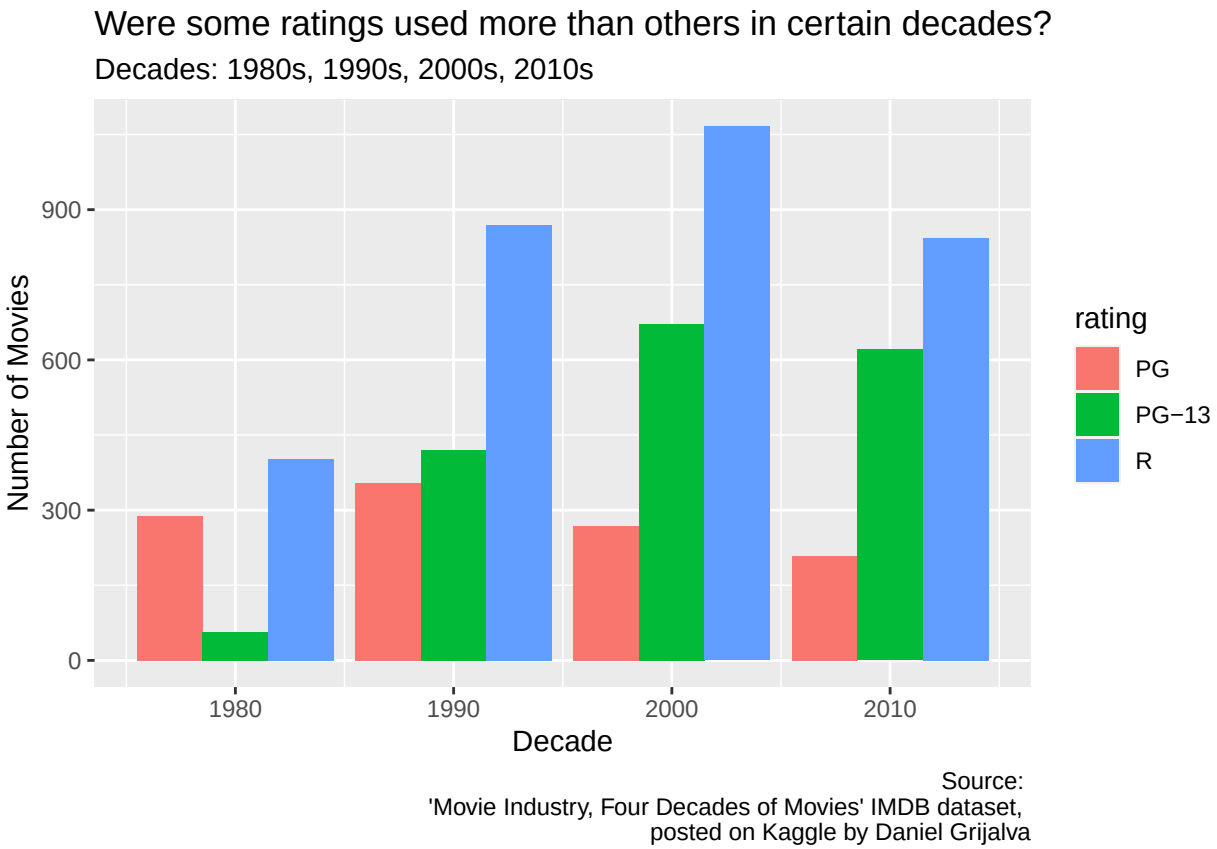
More horror movies were made in 00s-10s than in 80s-90s.

Were some ratings used more than others in certain decades?

```
ratingdecade <- movies %>%
  filter(!is.na(rating), decade != 2020) %>%
  filter(rating == "PG" | rating == "PG-13" | rating == "R") %>%
  group_by(decade) %>%
  count(rating, sort = TRUE) %>%
  arrange(rating)
```

Plot via clustered column chart:

```
ggplot(ratingdecade) +
  geom_col(aes(x = decade, y = n, fill = rating), position = "dodge") +
  labs(x = "Decade", y = "Number of Movies",
       title = "Were some ratings used more than others in certain decades?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



More PG rated movies were made in 80s-90s than in 00s-10s.

More PG-13 rated movies were made in 00s-10s than in 80s-90s (inverse relationship between PG and PG-13 in these decades).

More R rated movies were made in 90s-00s than in 80s/10s.

RECAP OF INSIGHTS

- Except for Avatar and Titanic, all the Top 20 grossing movies premiered in the last decade, and were mostly franchise-related.
- The Top 20 profitable (%) movies are more spread out over the decades, none are sequels, and no stars, directors, writers, or companies appear more than once.
- The most profitable (%) movies generally succeeded despite their low budget.
- None of the movies in the top 20 profit (%) movies had a gross above $370M.
- All of the movies in the top 20 grossing movies had a profit ($) above $1B.
- The most profitable (%) movies did not have the highest budgets or gross, but the larger budgets tended to create larger gross.
- There have been more R rated movies made than movies with any other rating. None of the top 20 grossing movies were rated R, but the majority of top 20 profitable (%) movies were rated R.
- The top 20 grossing movies were significantly longer compared to all movies with the same ratings; the top 20 profitable (%) movies were generally shorter.

- Action movies accounted for 13 out of the top 20 grossing films; Horror and Comedy combined for 13 out of the top 20 profitable (%) films.
- Marvel had the most profit of all film companies ($12B), followed by Dreamworks ($8B), Pixar ($6B), and Lucasfilm ($6B). Marvel made about $1B profit per movie.
- FilmDistrict, Exclusive Media Group, and Twisted Pictures all averaged over 1000% profit from their movies.
- There was a cumulative rise in gross among all movies each decade.
- Movies were significantly more profitable (%) in the 2010s (1157%), followed by 2000s (437%), 1980s (329%), and 1990s (222%).
- On average, movies get longer every decade.

PREPARE AND EXPORT CSV FOR TABLEAU

Revert the gross, budget, and profit to their original values for Tableau:

```
tabmovies <- movies
tabmovies$gross <- movies$grossM * 1000000
tabmovies$budget <- movies$budgetM * 1000000
tabmovies$profit <- movies$profitM * 1000000
tabmovies <- tabmovies %>%
  select(-c(grossM, budgetM, profitM))
```

Are there any duplicate movie titles?

```
length(unique(tabmovies$name)) == nrow(tabmovies)
```

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

There are some duplicate movie titles in this data set, which causes issues in Tableau. For example, Tableau will combine the gross of “The Lion King” (1994) and the "The Lion King (2019), thinking that these are the same movie. This skews the output of my dashboard.

Paste Name and Premiere together to make each movie name unique:

```
tabmovies$name <- paste(tabmovies$name, tabmovies$premiere, sep = " (")
```

Append a closing parenthesis:

```
tabmovies$name <- paste(tabmovies$name, ")", sep = "")
```

Export as CSV <- write.csv(movies, “filepath/filename.csv”, row.names = FALSE):