Movie Correlation and Analysis in R

Ed Garcia

8/29/2021

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.

- Since ticket sales were not part of this data set, I used the data from "Average ticket price at U.S. movie theaters 1980-2019" published on Statista.com by Jose Gabriel Navarro on Aug 12, 2021.
- 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(kableExtra) # for pretty tables
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
```

DATA CLEANING

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

Inspect the discrepancy:

```
kable(head(movies[, c(1,4,5)], 10)) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"), full_width = FALSE)
```

name	year	released
The Shining	1980	June 13, 1980 (United States)
The Blue Lagoon	1980	July 2, 1980 (United States)
Star Wars: Episode V - The Empire Strikes Back	1980	June 20, 1980 (United States)
Airplane!	1980	July 2, 1980 (United States)
Caddyshack	1980	July 25, 1980 (United States)
Friday the 13th	1980	May 9, 1980 (United States)
The Blues Brothers	1980	June 20, 1980 (United States)
Raging Bull	1980	December 19, 1980 (United States)
Superman II	1980	June 19, 1981 (United States)
The Long Riders	1980	May 16, 1980 (United States)

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 finances adjusted for inflation

Import Ticket Price Data: ticketprices <- read.csv("filepath/filename.csv")

Find 2019 ticket prices:

```
ticketprices %>%
  filter(Year == 2019)
```

```
##   Year AvgUSTicketPrice
## 1 2019             9.16
```

Add column for inflation adjustment to 2019 prices:

```
ticketprices$inflation <- 9.16 / ticketprices$AvgUSTicketPrice
```

Merge ticketprices data frame with movies data frame:

```
ticketprices <- ticketprices %>%
  rename(premiere = Year)
movies <- merge(movies, ticketprices, by = "premiere")
```

Calculate adjust finances to 2019 prices:

```
movies$adjbudget <- movies$budget * movies$inflation
movies$adjgross <- movies$gross * movies$inflation
```

Determine profit for each movie.

Make a Profit column:

```
movies$profit <- movies$gross - movies$budget
movies$adjprofit <- movies$adjgross - movies$adjbudget
```

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$adjgross <- movies$adjgross / 1000000
movies$adjbudget <- movies$adjbudget / 1000000
movies$adjprofit <- movies$adjprofit / 1000000
movies <- movies %>%
rename(grossM = gross, budgetM = budget, profitM = profit,
       adjgrossM = adjgross, adjbudgetM = adjbudget, adjprofitM = adjprofit)
```

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] "PG"        "R"         "G"         "Not Rated" ""          "NC-17"
##  [7] "TV-PG"     "Approved"  "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")
```

```
##     premiere               name   rating     genre score votes   director
## 204     1981 Tarzan the Ape Man Approved Adventure   3.4  5300 John Derek
##       writer     star       country budgetM   grossM                  company
## 204 Tom Rowe Bo Derek United States     6.5 36.56528 Metro-Goldwyn-Mayer (MGM)
##     runtime full_release_date full_release_location AvgUSTicketPrice inflation
## 204     115        1981-07-24         United States             2.78  3.294964
##     adjbudgetM adjgrossM  profitM adjprofitM profit_percent
## 204   21.41727  120.4813 30.06528   99.06402       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"
```

```r
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:

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

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

Unite Musical and Music ratings:

```r
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:

```r
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

I will investigate the variables involved in the highest grossing and most profitable ($ and %) movies from the perspective of the actual and inflation-adjusted finances.

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

```
movies %>%
  select(c(name, premiere, star, director, writer, company, grossM)) %>%
  arrange(desc(grossM)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Grossing Movies (actual): Personnel", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1,3:6), width = "3cm")
```

Table 1: Top 20 Grossing Movies (actual): Personnel

name	premiere	star	director	writer	company
Avatar	2009	Sam Worthington	James Cameron	James Cameron	Twentieth Century Fox
Avengers: Endgame	2019	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
Titanic	1997	Leonardo DiCaprio	James Cameron	James Cameron	Twentieth Century Fox
Star Wars: Episode VII - The Force Awakens	2015	Daisy Ridley	J.J. Abrams	Lawrence Kasdan	Lucasfilm
Avengers: Infinity War	2018	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
The Lion King	2019	Donald Glover	Jon Favreau	Jeff Nathanson	Walt Disney Pictures
Jurassic World	2015	Chris Pratt	Colin Trevorrow	Rick Jaffa	Universal Pictures
The Avengers	2012	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
Furious 7	2015	Vin Diesel	James Wan	Chris Morgan	Universal Pictures
Frozen II	2019	Kristen Bell	Chris Buck	Jennifer Lee	Walt Disney Animation Studios
Avengers: Age of Ultron	2015	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
Black Panther	2018	Chadwick Boseman	Ryan Coogler	Ryan Coogler	Marvel Studios
Harry Potter and the Deathly Hallows: Part 2	2011	Daniel Radcliffe	David Yates	Steve Kloves	Warner Bros.
Star Wars: Episode VIII - The Last Jedi	2017	Daisy Ridley	Rian Johnson	Rian Johnson	Walt Disney Pictures
Jurassic World: Fallen Kingdom	2018	Chris Pratt	J.A. Bayona	Derek Connolly	Universal Pictures
Frozen	2013	Kristen Bell	Chris Buck	Jennifer Lee	Walt Disney Animation Studios
Beauty and the Beast	2017	Emma Watson	Bill Condon	Stephen Chbosky	Mandeville Films
Incredibles 2	2018	Craig T. Nelson	Brad Bird	Brad Bird	Walt Disney Pictures
The Fate of the Furious	2017	Vin Diesel	F. Gary Gray	Gary Scott Thompson	Universal Pictures
Iron Man 3	2013	Robert Downey Jr.	Shane Black	Drew Pearce	Marvel Studios

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 highest grossing movies (adjusted for inflation)?

```
movies %>%
  select(c(name, premiere, star, director, writer, company, adjgrossM)) %>%
  arrange(desc(adjgrossM)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Grossing Movies (inflation-adjusted to 2019): Personnel", digits = 0) %
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1,3:6), width = "3cm")
```

Adjusting for inflation, the top 3 grossing movies were the same but in different order (Adjusted: 1.Titanic, 2.Avatar, 3.Avengers:Endgame vs. Actual: 1.Avatar, 2.Avengers:Endgame, 3.Titanic)

4.E.T., 5.The Lion King (1994), and 6.Jurassic Park were highly successful for their time. None were on the Top 20 Actual Grossing List.

None of movies in the top 6 inflated list were franchise-related at the time of their release except Avengers:Endgame. All the other 14 movies were franchise-related except Independence Day. The premiere years were much more spread out when compared to the Top 20 Actual Grossing List.

Still no female directors, but there were four female writers: Melissa Mathison (E.T.), Irene Mecchi (The Lion King), Leigh Brackett (Star Wars V: The Empire Strikes Back), and J.K. Rowling (Harry Potter and the Sorcerer's Stone).

Table 2: Top 20 Grossing Movies (inflation-adjusted to 2019): Personnel

name	premiere	star	director	writer	company
Titanic	1997	Leonardo DiCaprio	James Cameron	James Cameron	Twentieth Century Fox
Avatar	2009	Sam Worthington	James Cameron	James Cameron	Twentieth Century Fox
Avengers: Endgame	2019	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
E.T. the Extra-Terrestrial	1982	Henry Thomas	Steven Spielberg	Melissa Mathison	Universal Pictures
The Lion King	1994	Matthew Broderick	Roger Allers	Irene Mecchi	Walt Disney Pictures
Jurassic Park	1993	Sam Neill	Steven Spielberg	Michael Crichton	Universal Pictures
Star Wars: Episode VII - The Force Awakens	2015	Daisy Ridley	J.J. Abrams	Lawrence Kasdan	Lucasfilm
Avengers: Infinity War	2018	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
Star Wars: Episode I - The Phantom Menace	1999	Ewan McGregor	George Lucas	George Lucas	Lucasfilm
Star Wars: Episode V - The Empire Strikes Back	1980	Mark Hamill	Irvin Kershner	Leigh Brackett	Lucasfilm
Jurassic World	2015	Chris Pratt	Colin Trevorrow	Rick Jaffa	Universal Pictures
The Avengers	2012	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
The Lord of the Rings: The Return of the King	2003	Elijah Wood	Peter Jackson	J.R.R. Tolkien	New Line Cinema
Independence Day	1996	Will Smith	Roland Emmerich	Dean Devlin	Twentieth Century Fox
The Lion King	2019	Donald Glover	Jon Favreau	Jeff Nathanson	Walt Disney Pictures
Furious 7	2015	Vin Diesel	James Wan	Chris Morgan	Universal Pictures
Harry Potter and the Sorcerer's Stone	2001	Daniel Radcliffe	Chris Columbus	J.K. Rowling	Warner Bros.
Harry Potter and the Deathly Hallows: Part 2	2011	Daniel Radcliffe	David Yates	Steve Kloves	Warner Bros.
Avengers: Age of Ultron	2015	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
The Lord of the Rings: The Two Towers	2002	Elijah Wood	Peter Jackson	J.R.R. Tolkien	New Line Cinema

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

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

```
arrange(desc(profitM)) %>%
top_n(20) %>%
knitr::kable(caption = "Top 20 Profitable Movies (actual): Personnel", digits = 0) %>%
kableExtra::kable_styling(latex_options = c("hold_position")) %>%
column_spec(c(1,3:6), width = "3cm")
```

Table 3: Top 20 Profitable Movies (actual): Personnel

name	premiere	star	director	writer	company
Avatar	2009	Sam Worthington	James Cameron	James Cameron	Twentieth Century Fox
Avengers: Endgame	2019	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
Titanic	1997	Leonardo DiCaprio	James Cameron	James Cameron	Twentieth Century Fox
Star Wars: Episode VII - The Force Awakens	2015	Daisy Ridley	J.J. Abrams	Lawrence Kasdan	Lucasfilm
Avengers: Infinity War	2018	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
Jurassic World	2015	Chris Pratt	Colin Trevorrow	Rick Jaffa	Universal Pictures
The Lion King	2019	Donald Glover	Jon Favreau	Jeff Nathanson	Walt Disney Pictures
Furious 7	2015	Vin Diesel	James Wan	Chris Morgan	Universal Pictures
Frozen II	2019	Kristen Bell	Chris Buck	Jennifer Lee	Walt Disney Animation Studios
The Avengers	2012	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
Harry Potter and the Deathly Hallows: Part 2	2011	Daniel Radcliffe	David Yates	Steve Kloves	Warner Bros.
Avengers: Age of Ultron	2015	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
Black Panther	2018	Chadwick Boseman	Ryan Coogler	Ryan Coogler	Marvel Studios
Jurassic World: Fallen Kingdom	2018	Chris Pratt	J.A. Bayona	Derek Connolly	Universal Pictures
Frozen	2013	Kristen Bell	Chris Buck	Jennifer Lee	Walt Disney Animation Studios
Beauty and the Beast	2017	Emma Watson	Bill Condon	Stephen Chbosky	Mandeville Films
Minions	2015	Sandra Bullock	Kyle Balda	Brian Lynch	Illumination Entertainment
The Lord of the Rings: The Return of the King	2003	Elijah Wood	Peter Jackson	J.R.R. Tolkien	New Line Cinema
Incredibles 2	2018	Craig T. Nelson	Brad Bird	Brad Bird	Walt Disney Pictures
The Lion King	1994	Matthew Broderick	Roger Allers	Irene Mecchi	Walt Disney Pictures

Not much difference between the variables in the Top 20 profitable (actual $) and the Top 20 grossing (actual $) movies.

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

```
movies %>%
  select(c(name, premiere, star, director, writer, company, adjprofitM)) %>%
  arrange(desc(adjprofitM)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Profitable Movies (inflation-adjusted $): Personnel", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1,3:6), width = "3cm")
```

\begin{table}[!h]

\caption{Top 20 Profitable Movies (inflation-adjusted $): Personnel}

name	premiere	star	director	writer	company
Titanic	1997	Leonardo DiCaprio	James Cameron	James Cameron	Twentieth Century Fox
Avatar	2009	Sam Worthington	James Cameron	James Cameron	Twentieth Century Fox
Avengers: Endgame	2019	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
E.T. the Extra-Terrestrial	1982	Henry Thomas	Steven Spielberg	Melissa Mathison	Universal Pictures
The Lion King	1994	Matthew Broderick	Roger Allers	Irene Mecchi	Walt Disney Pictures
Jurassic Park	1993	Sam Neill	Steven Spielberg	Michael Crichton	Universal Pictures
Star Wars: Episode VII - The Force Awakens	2015	Daisy Ridley	J.J. Abrams	Lawrence Kasdan	Lucasfilm
Star Wars: Episode V - The Empire Strikes Back	1980	Mark Hamill	Irvin Kershner	Leigh Brackett	Lucasfilm
Avengers: Infinity War	2018	Robert Downey Jr.	Anthony Russo	Christopher Markus	Marvel Studios
Jurassic World	2015	Chris Pratt	Colin Trevorrow	Rick Jaffa	Universal Pictures
Star Wars: Episode I - The Phantom Menace	1999	Ewan McGregor	George Lucas	George Lucas	Lucasfilm
The Lord of the Rings: The Return of the King	2003	Elijah Wood	Peter Jackson	J.R.R. Tolkien	New Line Cinema
Independence Day	1996	Will Smith	Roland Emmerich	Dean Devlin	Twentieth Century Fox
The Avengers	2012	Robert Downey Jr.	Joss Whedon	Joss Whedon	Marvel Studios
Furious 7	2015	Vin Diesel	James Wan	Chris Morgan	Universal Pictures
Harry Potter and the Sorcerer's Stone	2001	Daniel Radcliffe	Chris Columbus	J.K. Rowling	Warner Bros.
The Lion King	2019	Donald Glover	Jon Favreau	Jeff Nathanson	Walt Disney Pictures
Harry Potter and the Deathly Hallows: Part 2	2011	Daniel Radcliffe	David Yates	Steve Kloves	Warner Bros.
Forrest Gump	1994	Tom Hanks	Robert Zemeckis	Winston Groom	Paramount Pictures
The Lord of the Rings: The Two Towers	2002	Elijah Wood	Peter Jackson	J.R.R. Tolkien	New Line Cinema

\end{table} The Top 20 Profitable (inflation-adjusted $) Movies list is very similar to the the Top 20 Grossing (inflation-adjusted $) list.

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, profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Profit Percentage Movies: Personnel", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1,3:6), width = "3cm")
```

Table 4: Top 20 Profit Percentage Movies: Personnel

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

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.
- In general, these movies succeeded in profitability despite their low budget.

Look at the financial numbers of these 3 lists (actual):

```r
movies %>%
  select(c(name, budgetM, profitM, profit_percent, grossM)) %>%
  arrange(desc(grossM)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Grossing Movies (actual): Finances", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

Table 5: Top 20 Grossing Movies (actual): Finances

name	budgetM	profitM	profit_percent	grossM
Avatar	237	2610	1101	2847
Avengers: Endgame	356	2442	686	2798
Titanic	200	2002	1001	2202
Star Wars: Episode VII - The Force Awakens	245	1825	745	2070
Avengers: Infinity War	321	1727	538	2048
The Lion King	260	1411	543	1671
Jurassic World	150	1521	1014	1671
The Avengers	220	1299	590	1519
Furious 7	190	1325	698	1515
Frozen II	150	1300	867	1450
Avengers: Age of Ultron	250	1153	461	1403
Black Panther	200	1148	574	1348
Harry Potter and the Deathly Hallows: Part 2	125	1217	974	1342
Star Wars: Episode VIII - The Last Jedi	317	1016	320	1333
Jurassic World: Fallen Kingdom	170	1140	671	1310
Frozen	150	1132	754	1282
Beauty and the Beast	160	1104	690	1264
Incredibles 2	200	1045	522	1245
The Fate of the Furious	250	986	394	1236
Iron Man 3	200	1015	507	1215

```r
movies %>%
  select(c(name, budgetM, grossM, profit_percent, profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Profitable Movies (actual): Finances", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

Table 6: Top 20 Profitable Movies (actual): Finances

name	budgetM	grossM	profit_percent	profitM
Avatar	237	2847	1101	2610
Avengers: Endgame	356	2798	686	2442
Titanic	200	2202	1001	2002
Star Wars: Episode VII - The Force Awakens	245	2070	745	1825
Avengers: Infinity War	321	2048	538	1727
Jurassic World	150	1671	1014	1521
The Lion King	260	1671	543	1411
Furious 7	190	1515	698	1325
Frozen II	150	1450	867	1300
The Avengers	220	1519	590	1299
Harry Potter and the Deathly Hallows: Part 2	125	1342	974	1217
Avengers: Age of Ultron	250	1403	461	1153
Black Panther	200	1348	574	1148
Jurassic World: Fallen Kingdom	170	1310	671	1140
Frozen	150	1282	754	1132
Beauty and the Beast	160	1264	690	1104
Minions	74	1159	1467	1085
The Lord of the Rings: The Return of the King	94	1146	1119	1052
Incredibles 2	200	1245	522	1045
The Lion King	45	1084	2308	1039

```r
movies %>%
  select(c(name, budgetM, grossM, profitM, profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(20) %>%
  knitr::kable(caption = "Top 20 Profit Percentage Movies: Finances", digits = 3) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

Table 7: Top 20 Profit Percentage Movies: Finances

name	budgetM	grossM	profitM	profit_percent
Paranormal Activity	0.015	193.356	193.341	1288938.667
The Blair Witch Project	0.060	248.639	248.579	414298.498
The Gallows	0.100	42.964	42.864	42864.410
El Mariachi	0.007	2.041	2.034	29056.000
Once	0.150	20.937	20.787	13857.815
Clerks	0.027	3.151	3.124	11570.852
Napoleon Dynamite	0.400	46.139	45.739	11434.722
In the Company of Men	0.025	2.804	2.779	11117.892
Keeping Mum	0.169	18.587	18.418	10898.127
Open Water	0.500	54.683	54.183	10836.697
The Devil Inside	1.000	101.758	100.758	10075.849
The Quiet Ones	0.200	17.835	17.635	8817.581
Saw	1.200	103.912	102.712	8559.306
Searching	0.880	75.462	74.582	8475.231
Primer	0.007	0.545	0.538	7691.943
E.T. the Extra-Terrestrial	10.500	792.911	782.411	7451.529
My Big Fat Greek Wedding	5.000	368.744	363.744	7274.881
The Full Monty	3.500	257.939	254.439	7269.676
Friday the 13th	0.550	39.755	39.205	7128.109
Fireproof	0.500	33.473	32.973	6594.659

Insights: Only 5 movies in the top 20 profit (%) movies had budgets of >= $1M, and none are in the top 10.

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 an actual gross above $370M.

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

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

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

CORRELATION MATRIX

Make correlation matrix for all variables:

```r
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 labels were applied correctly
```

```
##   premiere name rating genre score votes director writer star country budgetM
## 1     1980    0      0     0   6.9 29000        0      0    0       0      10
## 2     1980    1      1     1   6.8 66000        1      1    1       0       1
## 3     1980    2      1     0   6.8 26000        2      2    2       0      NA
## 4     1980    3      1     2   6.5 20000        3      3    3       1      11
## 5     1980    4      1     3   6.8 14000        4      4    4       0      44
## 6     1980    5      0     0   7.1  8900        5      5    5       0      NA
##       grossM company runtime full_release_date full_release_location
## 1 103.300686       0     109        1980-12-19                     0
## 2  21.448782       1      89        1980-02-08                     0
## 3 101.300000       2     111        1980-12-12                     0
## 4  19.814523       3     102        1980-02-15                     0
## 5   3.484523       4     219        1981-04-24                     0
## 6  22.482952       5     102        1980-09-26                     0
##   AvgUSTicketPrice inflation adjbudgetM adjgrossM    profitM adjprofitM
## 1             2.69  3.405204  34.052045 351.75996  93.300686  317.70791
## 2             2.69  3.405204   3.405204  73.03749  20.448782   69.63228
## 3             2.69  3.405204         NA 344.94721         NA         NA
## 4             2.69  3.405204  37.457249  67.47250   8.814523   30.01525
## 5             2.69  3.405204 149.828996  11.86551 -40.515477 -137.96348
## 6             2.69  3.405204         NA  76.55905         NA         NA
##   profit_percent decade
## 1      933.00686   1980
## 2     2044.87820   1980
## 3             NA   1980
## 4       80.13203   1980
## 5      -92.08063   1980
## 6             NA   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)
  {
```

```r
  corr_matrix[lower.tri(corr_matrix)]<- NA
  return(corr_matrix)
  }
```

Return usable data frame:

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

```
##                       premiere name rating genre score votes director writer
## premiere                     1 0.95   0.20  0.13  0.05  0.21     0.73   0.78
## name                        NA 1.00   0.19  0.13  0.06  0.19     0.70   0.76
## rating                      NA   NA   1.00 -0.05 -0.04  0.10     0.11   0.13
## genre                       NA   NA     NA  1.00  0.20  0.13     0.08   0.11
## 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
## AvgUSTicketPrice            NA   NA     NA    NA    NA    NA       NA     NA
## inflation                   NA   NA     NA    NA    NA    NA       NA     NA
## adjbudgetM                  NA   NA     NA    NA    NA    NA       NA     NA
## adjgrossM                   NA   NA     NA    NA    NA    NA       NA     NA
## profitM                     NA   NA     NA    NA    NA    NA       NA     NA
## adjprofitM                  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
## premiere              0.71    0.09    0.32   0.27    0.52    0.07
## name                  0.68    0.09    0.30   0.24    0.50    0.06
## rating                0.13   -0.01    0.21   0.16    0.00    0.09
## genre                 0.08    0.10    0.22   0.18    0.06    0.14
## score                 0.00    0.09    0.07   0.22    0.04    0.41
## votes                 0.09   -0.02    0.44   0.62   -0.05    0.35
## director              0.63    0.07    0.09   0.14    0.45   -0.13
## writer                0.60    0.08    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   -0.04  -0.04    0.13    0.09
## budgetM                 NA      NA    1.00   0.74   -0.15    0.32
## grossM                  NA      NA      NA   1.00   -0.09    0.27
## 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
## AvgUSTicketPrice        NA      NA      NA     NA      NA      NA
## inflation               NA      NA      NA     NA      NA      NA
## adjbudgetM              NA      NA      NA     NA      NA      NA
## adjgrossM               NA      NA      NA     NA      NA      NA
## profitM                 NA      NA      NA     NA      NA      NA
```

```
## adjprofitM              NA      NA      NA     NA      NA      NA
## profit_percent          NA      NA      NA     NA      NA      NA
## decade                  NA      NA      NA     NA      NA      NA
##                      full_release_location AvgUSTicketPrice inflation
## premiere                              0.15             0.98     -0.97
## name                                  0.15             0.93     -0.93
## rating                                0.01             0.19     -0.22
## genre                                -0.01             0.14     -0.12
## score                                 0.00             0.06     -0.05
## votes                                -0.05             0.20     -0.20
## director                              0.13             0.73     -0.71
## writer                                0.13             0.77     -0.77
## star                                  0.12             0.70     -0.69
## country                               0.06             0.09     -0.09
## budgetM                              -0.06             0.31     -0.32
## grossM                               -0.05             0.27     -0.25
## company                               0.16             0.51     -0.49
## runtime                              -0.04             0.07     -0.06
## full_release_location                 1.00             0.16     -0.15
## AvgUSTicketPrice                        NA             1.00     -0.95
## inflation                               NA               NA      1.00
## adjbudgetM                              NA               NA        NA
## adjgrossM                               NA               NA        NA
## profitM                                 NA               NA        NA
## adjprofitM                              NA               NA        NA
## profit_percent                          NA               NA        NA
## decade                                  NA               NA        NA
##                      adjbudgetM adjgrossM profitM adjprofitM profit_percent
## premiere                   0.12      0.13    0.24       0.12           0.01
## name                       0.10      0.11    0.21       0.10           0.01
## rating                     0.18      0.13    0.13       0.11          -0.01
## genre                      0.21      0.17    0.16       0.14          -0.02
## score                      0.05      0.24    0.24       0.27           0.00
## votes                      0.41      0.63    0.61       0.63           0.02
## director                  -0.08      0.03    0.14       0.05           0.02
## writer                     0.02      0.04    0.13       0.05           0.02
## star                      -0.06      0.03    0.13       0.05           0.02
## country                   -0.06     -0.05   -0.04      -0.05           0.00
## budgetM                    0.94      0.66    0.61       0.53          -0.02
## grossM                     0.66      0.95    0.98       0.92           0.02
## company                   -0.26     -0.16   -0.07      -0.12           0.02
## runtime                    0.34      0.28    0.24       0.24          -0.02
## full_release_location     -0.10     -0.07   -0.05      -0.06           0.00
## AvgUSTicketPrice           0.09      0.13    0.24       0.12           0.01
## inflation                 -0.12     -0.12   -0.22      -0.11          -0.01
## adjbudgetM                 1.00      0.66    0.53       0.51          -0.02
## adjgrossM                    NA      1.00    0.94       0.98           0.02
## profitM                      NA        NA    1.00       0.95           0.02
## adjprofitM                   NA        NA      NA       1.00           0.03
## profit_percent               NA        NA      NA         NA           1.00
## decade                       NA        NA      NA         NA             NA
##                      decade
## premiere               0.97
## name                   0.92
```

```
## rating                    0.18
## genre                     0.13
## score                     0.06
## votes                     0.18
## director                  0.71
## writer                    0.76
## star                      0.69
## country                   0.09
## budgetM                   0.30
## grossM                    0.26
## company                   0.51
## runtime                   0.07
## full_release_location     0.13
## AvgUSTicketPrice          0.95
## inflation                -0.94
## adjbudgetM                0.10
## adjgrossM                 0.12
## profitM                   0.22
## adjprofitM                0.11
## profit_percent            0.01
## decade                    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)
```

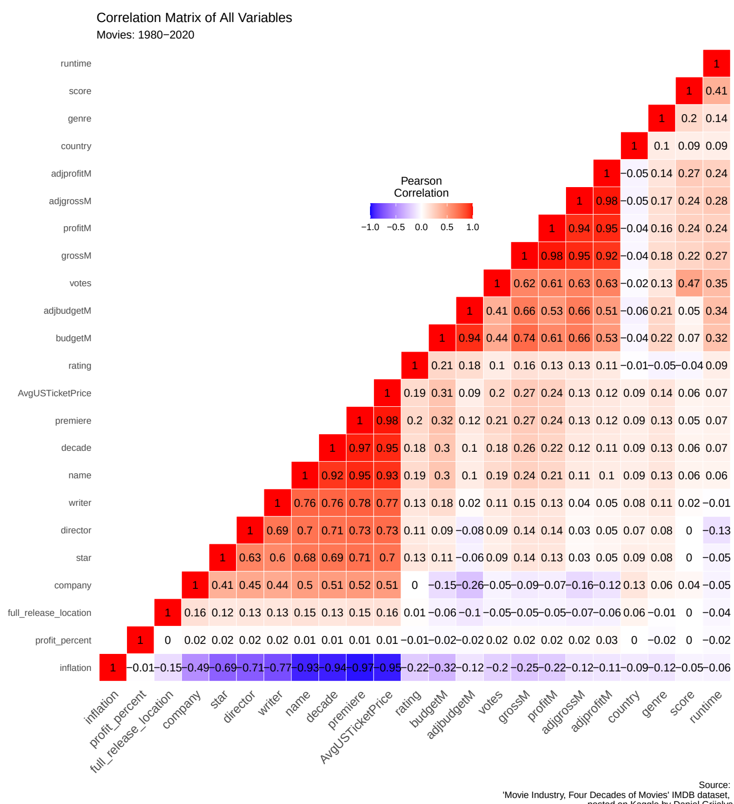
Plot the heat map:

```
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))
```



These correlations are highly logical, such as year-related variables (decades and premieres), inflation variables (actual vs. adjusted), 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:

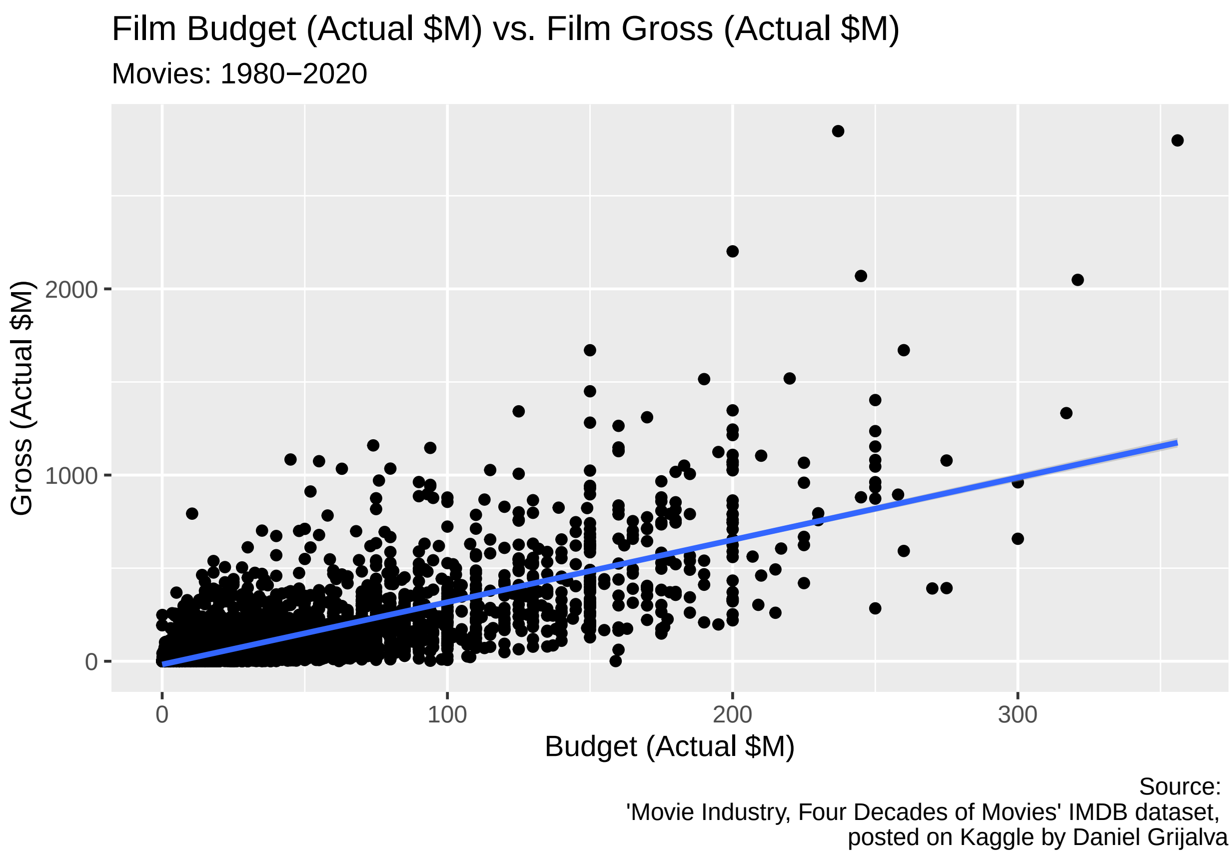
```
gross_high_corr <- melted_corr_matrix %>%
  filter(melted_corr_matrix$Var2 == "grossM"
         & melted_corr_matrix$value >= 0.5 # this only pulls highly correlated variables
         & melted_corr_matrix$value != 1) %>% # this ignores self-correlated variables
  arrange(desc(value))
tibble(gross_high_corr)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

```
## # A tibble: 3 x 3
##   Var1       Var2   value
##   <fct>      <fct>  <dbl>
## 1 budgetM    grossM  0.74
## 2 adjbudgetM grossM  0.66
## 3 votes      grossM  0.62
```

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:

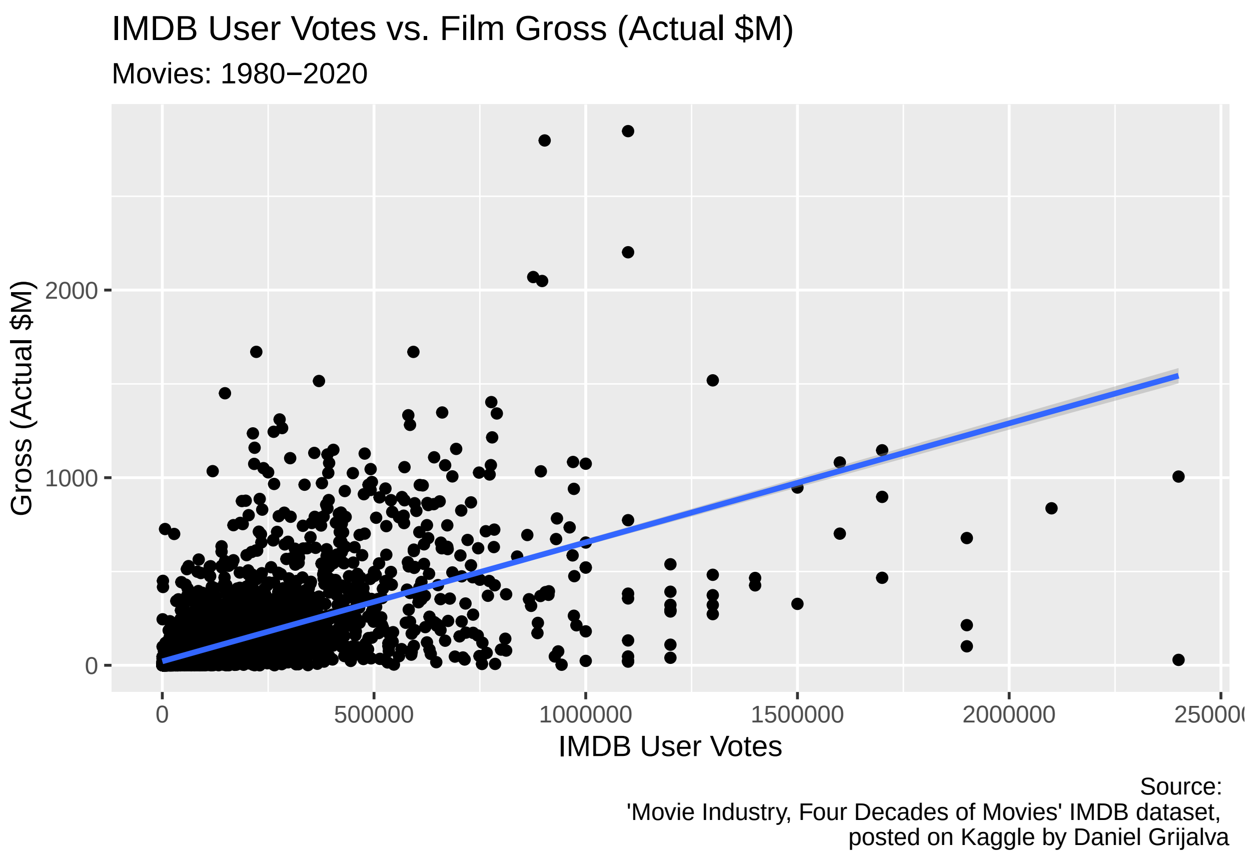
```
# the plot will skip over any NULLs
ggplot(movies, aes(budgetM, grossM)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Budget (Actual $M)", y = "Gross (Actual $M)",
       title = "Film Budget (Actual $M) vs. Film Gross (Actual $M)",
       subtitle = "Movies: 1980-2020",
       caption = "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 (Actual $M)",
       title = "IMDB User Votes vs. Film Gross (Actual $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: 4 x 3
##   Var1       Var2    value
##   <fct>      <fct>   <dbl>
## 1 grossM     profitM  0.98
## 2 budgetM    profitM  0.61
## 3 votes      profitM  0.61
## 4 adjbudgetM profitM  0.53
```

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))
```

```r
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.

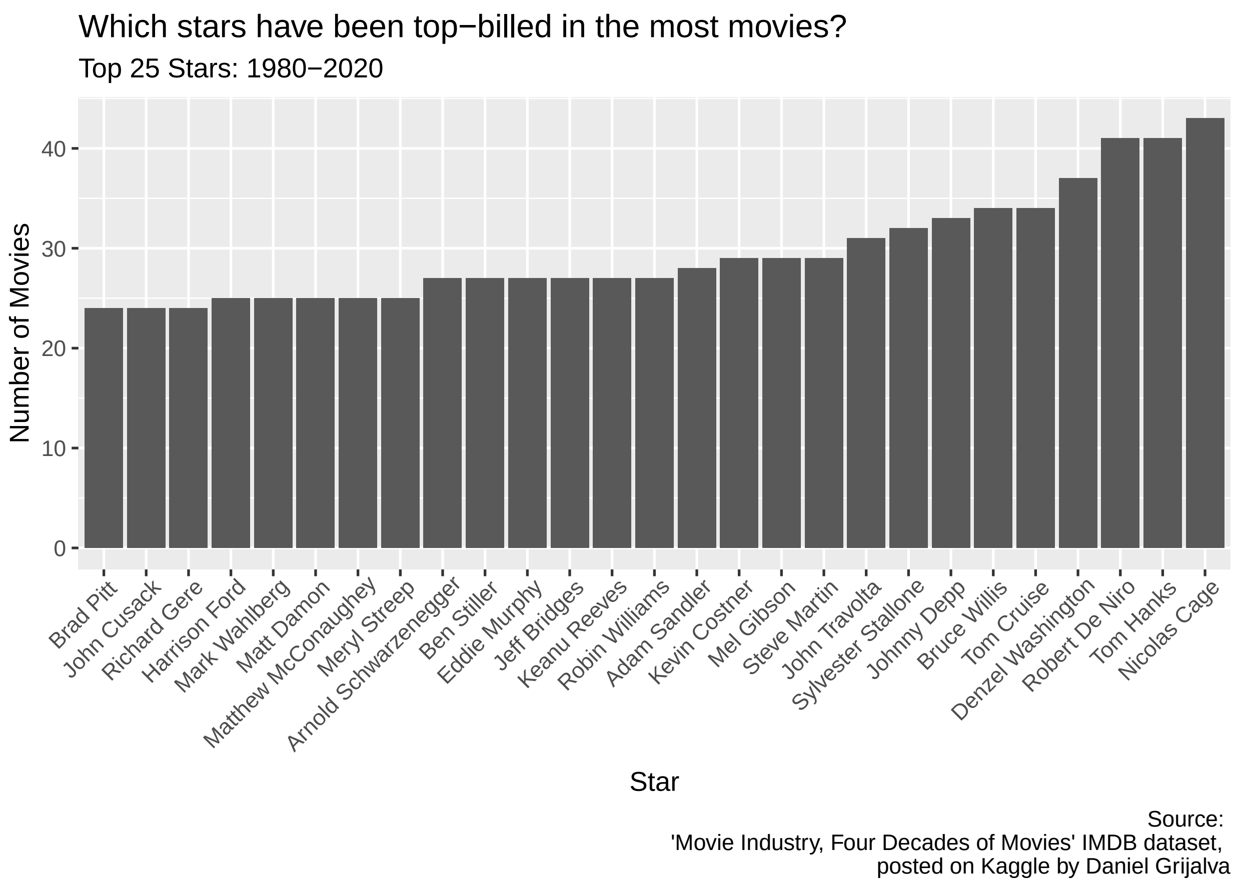
Total Range Analysis of Select Variables

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

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

Plot via column chart:

```r
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.

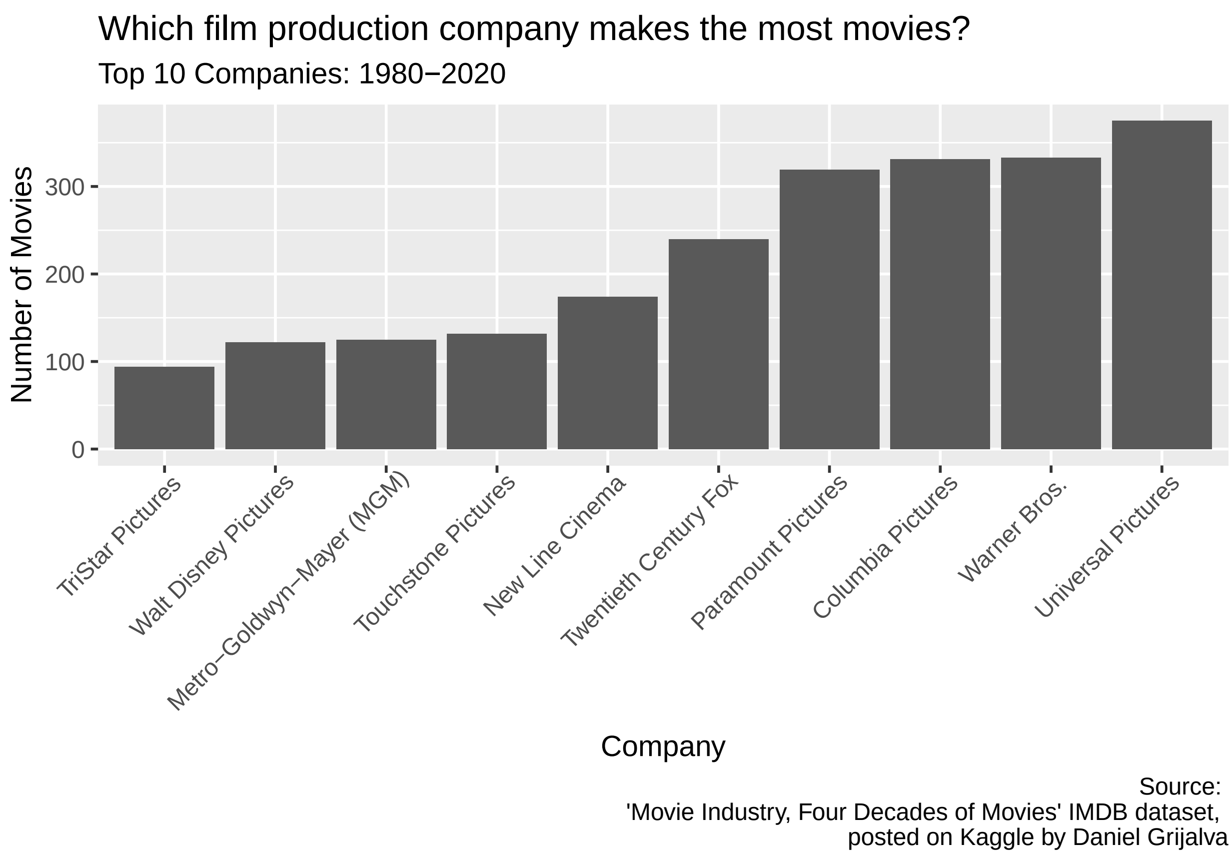
Note: I have to go all the way down to #24 before finding a female top-billed actor (Streep).

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.

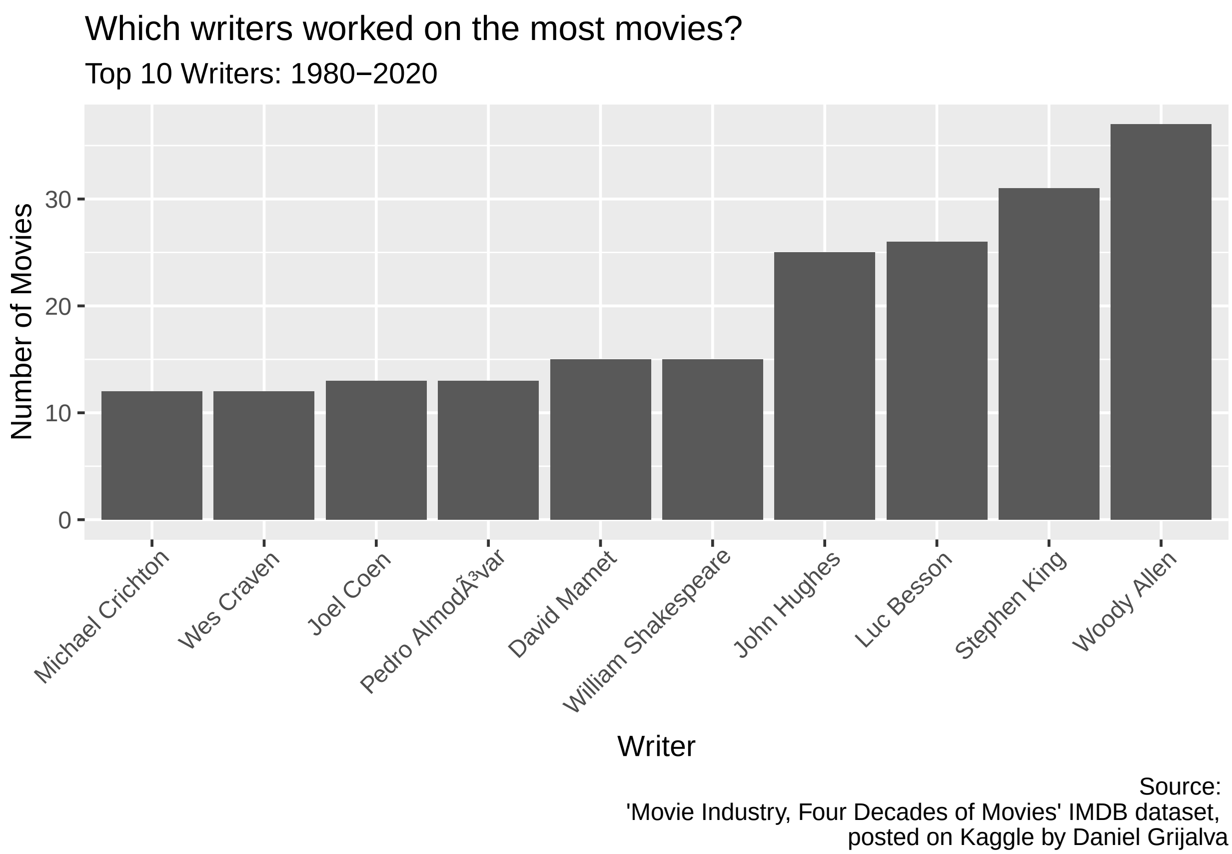
- Columbia and Paramount are not on the Top 20 grossing list.
- 20th Century and Walt Disney Pictures appear multiple times on the Top 20 grossing list, but made far less movies in the last 4 decades than the top 4 movie-making companies.

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.

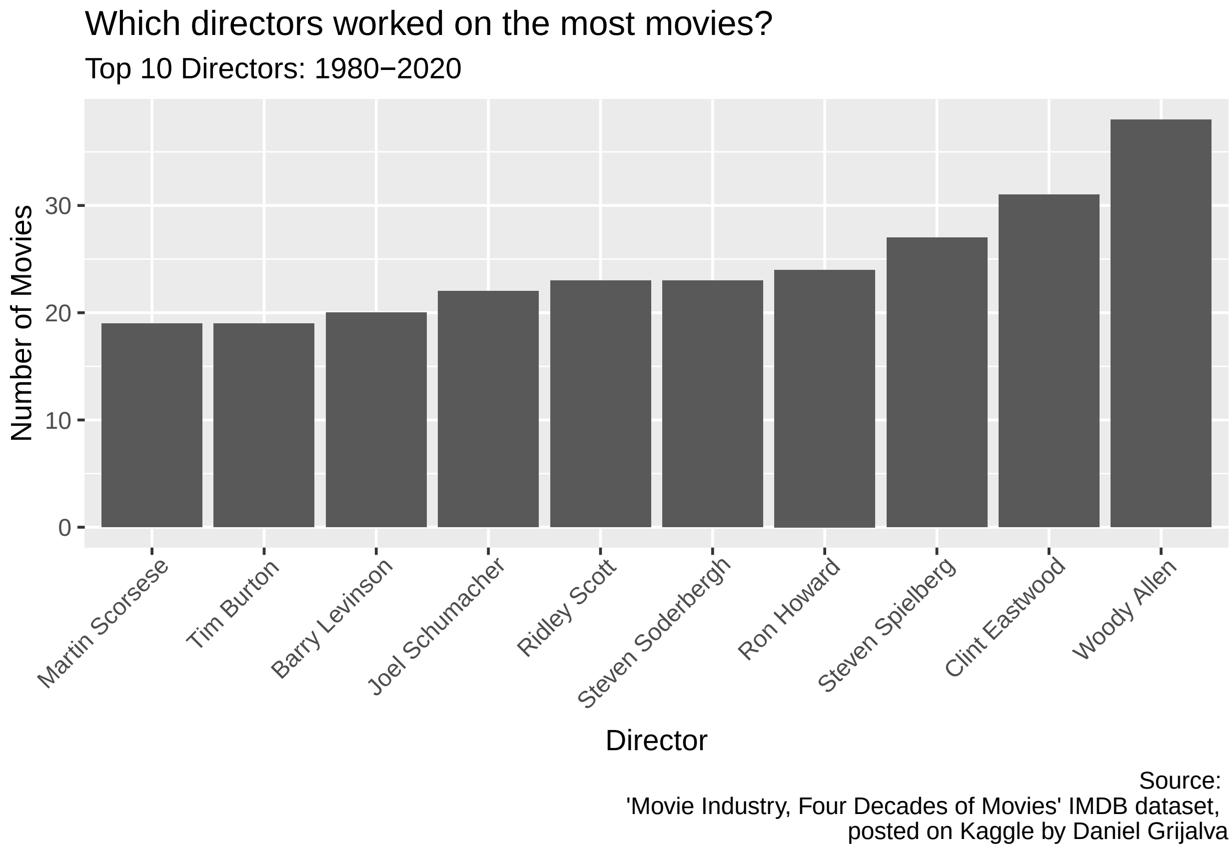
- None of the top 10 writers wrote the Top 20 grossing movies, nor the most profitable movies ($ or %).

Which directors worked on the most movies?

```
directors <- movies %>%
  count(director, sort = TRUE) %>%
  filter(director != "Directors") %>% #remove the co-directors value for clarity
  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.

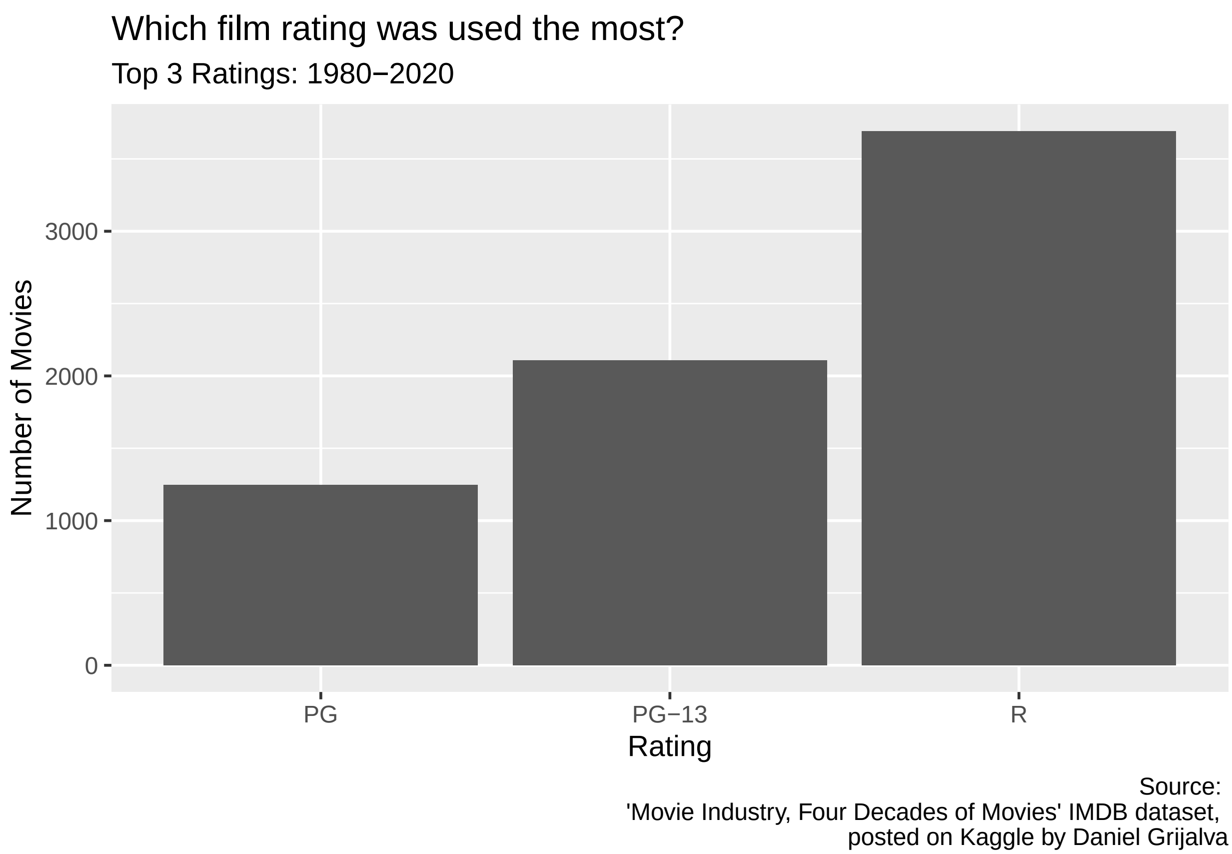
- None of the Top 10 directors wrote the Top 20 grossing movies, nor the most profitable movies ($ or %).

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")
```



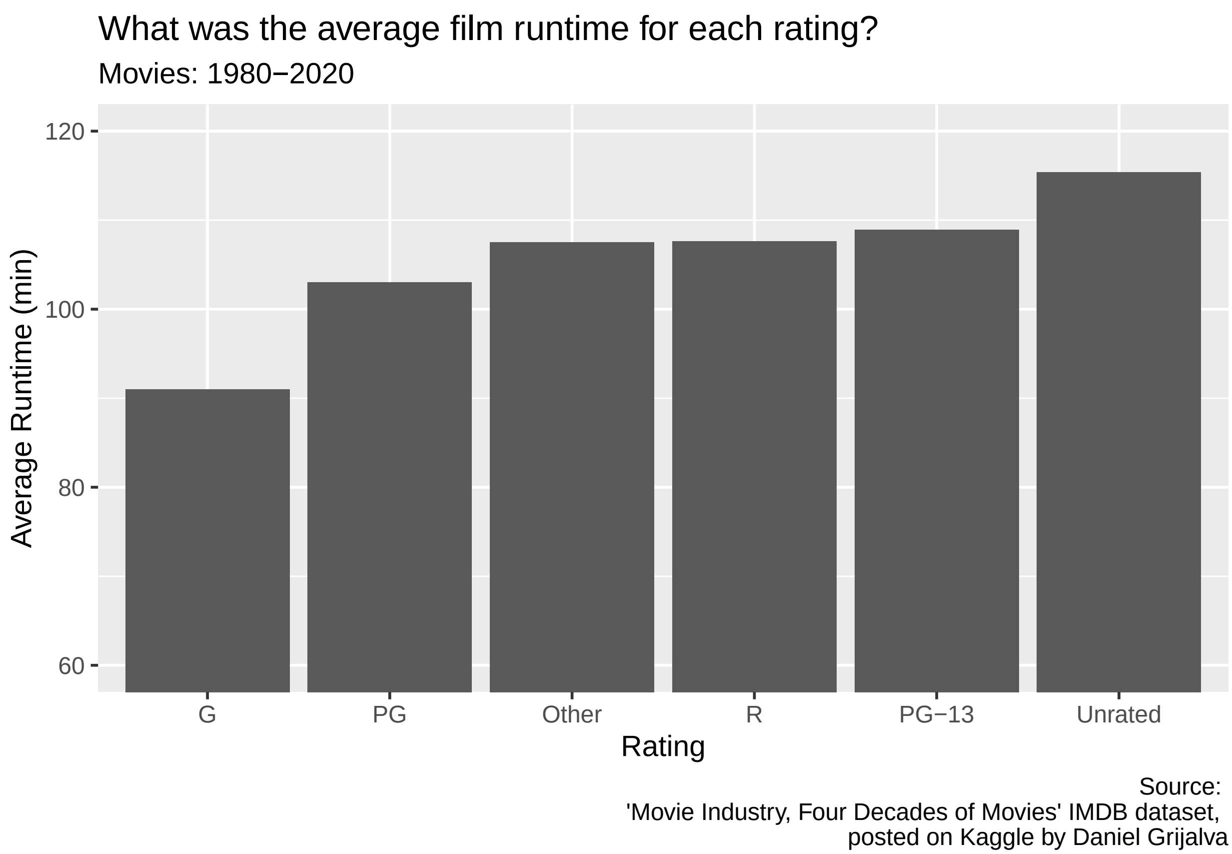
R rated films were made the most often (3698), followed by PG-13 (2112) and PG (1252).

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.

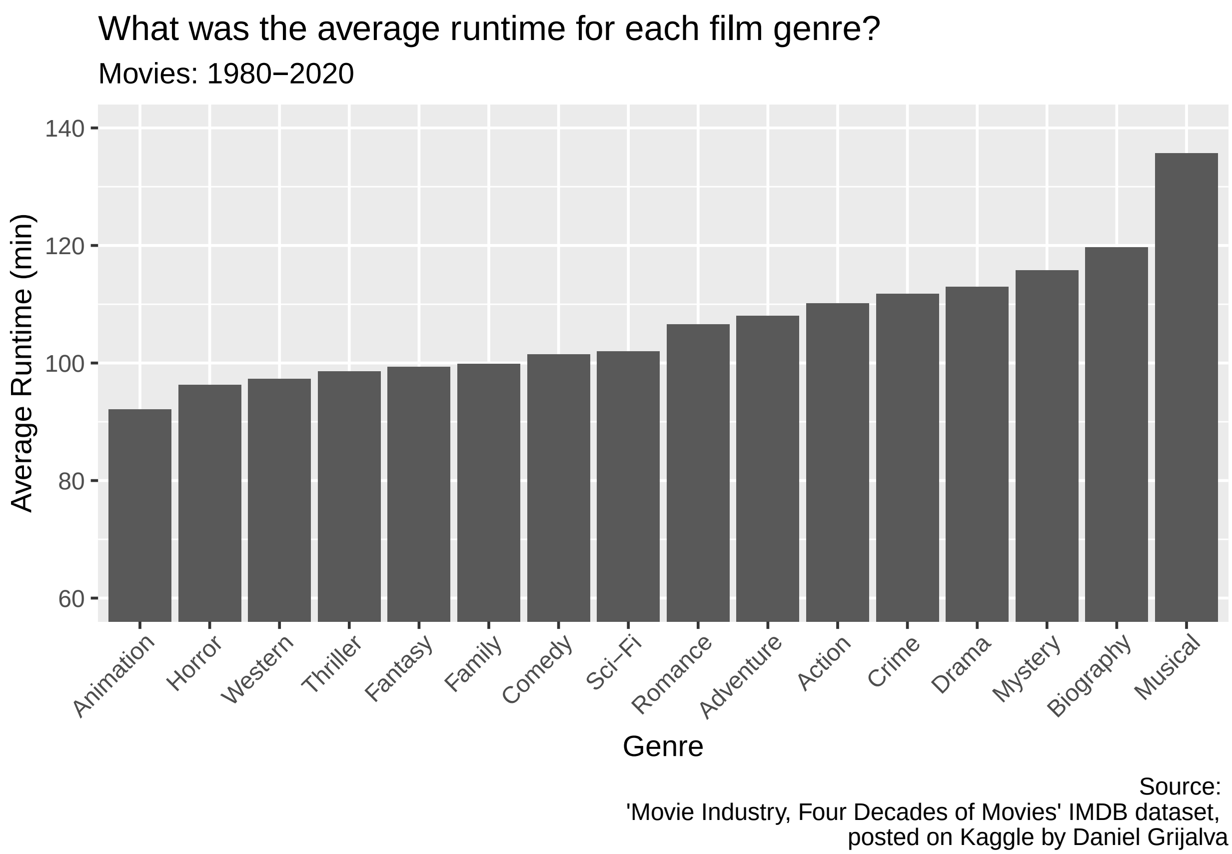
- PG and G make sense, since these are geared towards young children with shorter attention spans.

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?

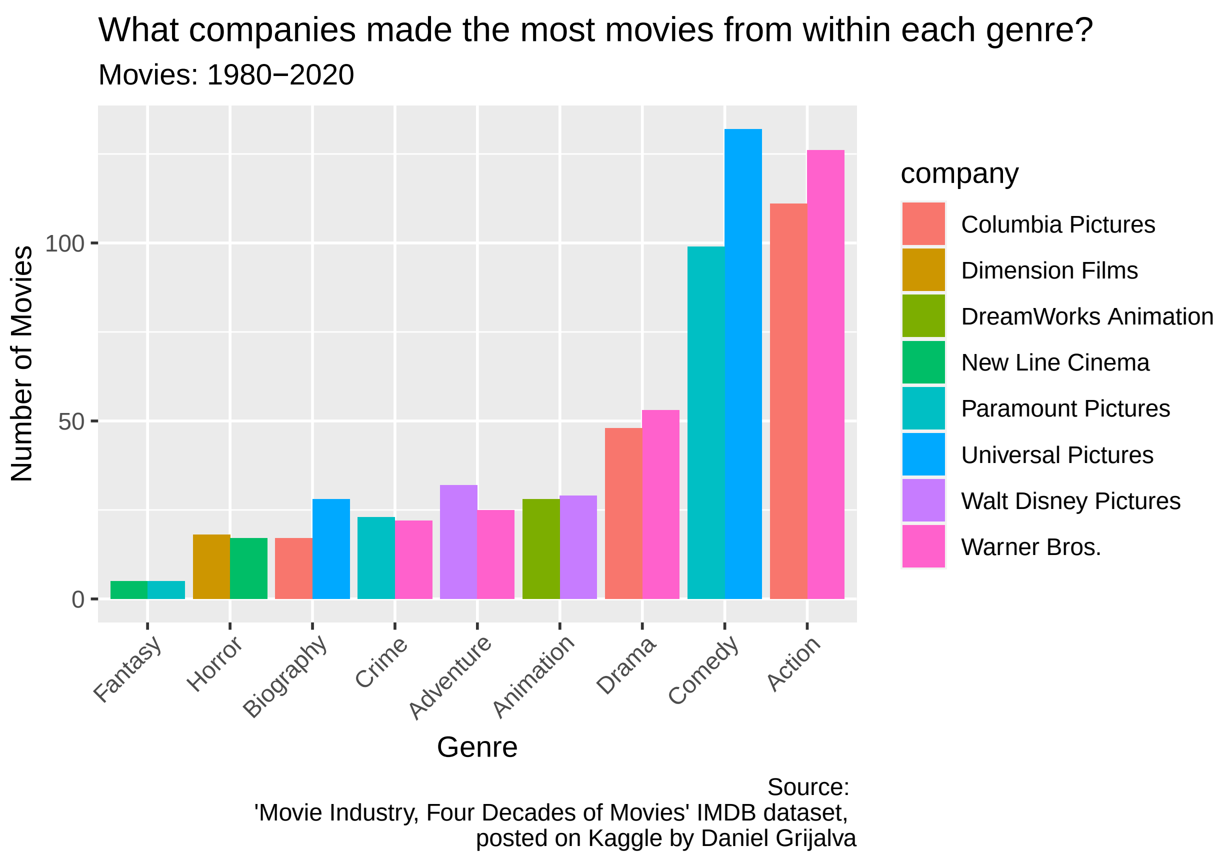
```
compGenre <- movies %>%
  group_by(genre, company) %>%
  count(genre, sort = TRUE)
compGenre <- compGenre %>%
  arrange(desc(n)) %>%
  group_by(genre) %>%
  slice(1:2)  # Top 2 highest values (number of movies made by company) by group (genre)
compGenre <- compGenre %>%  #some genres had < 10 movies, 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.           126
##  2 Action    Columbia Pictures      111
##  3 Adventure Walt Disney Pictures    32
##  4 Adventure Warner Bros.            25
##  5 Animation Walt Disney Pictures    29
##  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

I will use BLUE charts to represent the actual financial information, and ORANGE charts to represent the inflation-adjusted financial information.

Top 20 Highest Grossing Movies (actual):

```
top20gross <- movies %>%
  select(c(name, rating, runtime, genre, profitM, profit_percent, grossM)) %>%
  arrange(desc(grossM)) %>%
  top_n(20)
top20gross %>%
  knitr::kable(caption = "Top 20 Grossing Movies (actual): Categorical", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1), width = "3cm")
```

Table 8: Top 20 Grossing Movies (actual): Categorical

name	rating	runtime	genre	profitM	profit_percent	grossM
Avatar	PG-13	162	Action	2610	1101	2847
Avengers: Endgame	PG-13	181	Action	2442	686	2798
Titanic	PG-13	194	Drama	2002	1001	2202
Star Wars: Episode VII - The Force Awakens	PG-13	138	Action	1825	745	2070
Avengers: Infinity War	PG-13	149	Action	1727	538	2048
The Lion King	PG	118	Animation	1411	543	1671
Jurassic World	PG-13	124	Action	1521	1014	1671
The Avengers	PG-13	143	Action	1299	590	1519
Furious 7	PG-13	137	Action	1325	698	1515
Frozen II	PG	103	Animation	1300	867	1450
Avengers: Age of Ultron	PG-13	141	Action	1153	461	1403
Black Panther	PG-13	134	Action	1148	574	1348
Harry Potter and the Deathly Hallows: Part 2	PG-13	130	Adventure	1217	974	1342
Star Wars: Episode VIII - The Last Jedi	PG-13	152	Action	1016	320	1333
Jurassic World: Fallen Kingdom	PG-13	128	Action	1140	671	1310
Frozen	PG	102	Animation	1132	754	1282
Beauty and the Beast	PG	129	Family	1104	690	1264
Incredibles 2	PG	118	Animation	1045	522	1245
The Fate of the Furious	PG-13	136	Action	986	394	1236
Iron Man 3	PG-13	130	Action	1015	507	1215

Top 20 Highest Grossing Movies (inflation-adjusted):

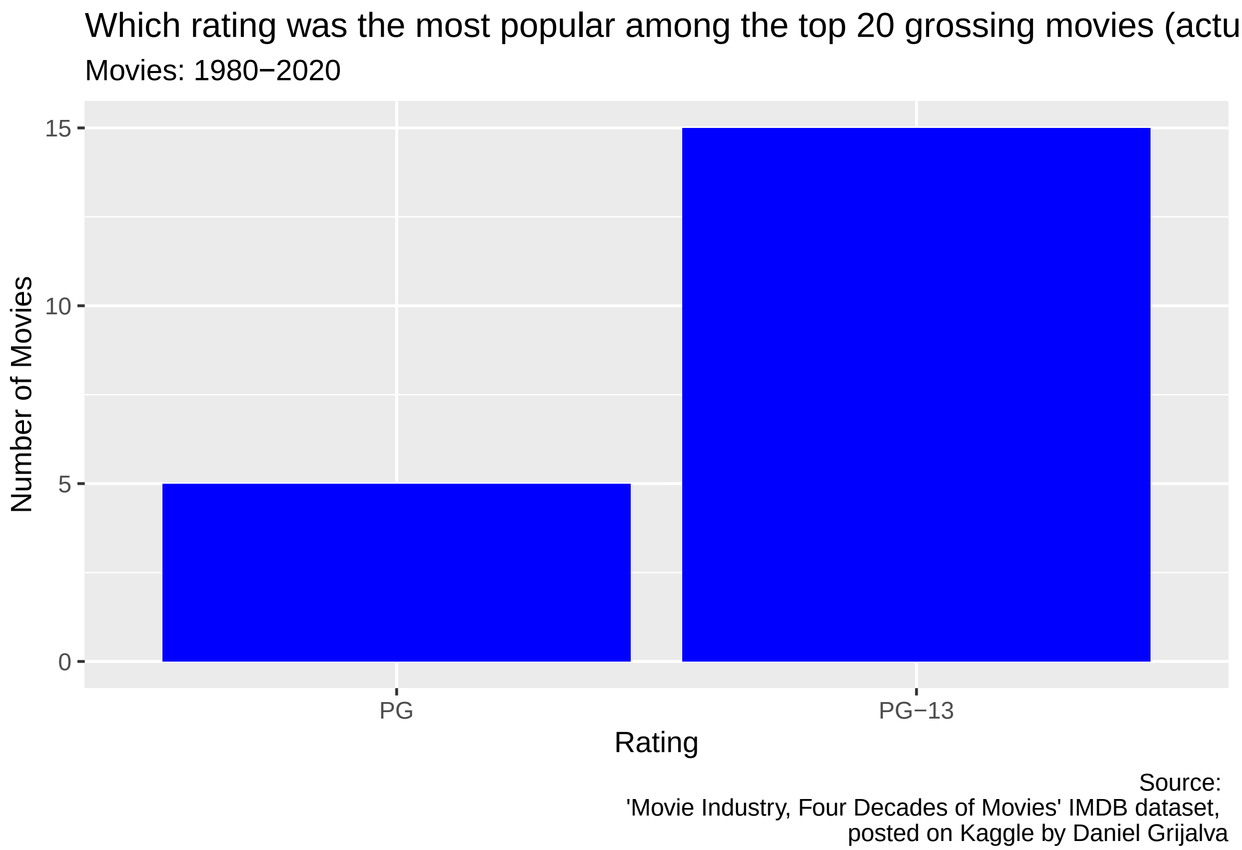
```r
adjtop20gross <- movies %>%
  select(c(name, rating, runtime, genre, adjprofitM, profit_percent, adjgrossM)) %>%
  arrange(desc(adjgrossM)) %>%
  top_n(20)
adjtop20gross %>%
  knitr::kable(caption = "Top 20 Grossing Movies (actual): Categorical", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position")) %>%
  column_spec(c(1), width = "3cm")
```

Table 9: Top 20 Grossing Movies (actual): Categorical

name	rating	runtime	genre	adjprofitM	profit_percent	adjgrossM
Titanic	PG-13	194	Drama	3995	1001	4394
Avatar	PG-13	162	Action	3188	1101	3477
Avengers: Endgame	PG-13	181	Action	2442	686	2798
E.T. the Extra-Terrestrial	PG	115	Family	2438	7452	2470
The Lion King	G	88	Animation	2276	2308	2375
Jurassic Park	PG-13	127	Action	2148	1541	2288
Star Wars: Episode VII - The Force Awakens	PG-13	138	Action	1983	745	2249
Avengers: Infinity War	PG-13	149	Action	1737	538	2060
Star Wars: Episode I - The Phantom Menace	PG	136	Action	1645	793	1852
Star Wars: Episode V - The Empire Strikes Back	PG	124	Action	1772	2891	1833
Jurassic World	PG-13	124	Action	1652	1014	1815
The Avengers	PG-13	143	Action	1495	590	1748
The Lord of the Rings: The Return of the King	PG-13	201	Action	1598	1119	1741
Independence Day	PG-13	145	Action	1539	990	1694
The Lion King	PG	118	Animation	1411	543	1671
Furious 7	PG-13	137	Action	1440	698	1647
Harry Potter and the Sorcerer's Stone	PG	152	Adventure	1427	706	1630
Harry Potter and the Deathly Hallows: Part 2	PG-13	130	Adventure	1406	974	1551
Avengers: Age of Ultron	PG-13	141	Action	1253	461	1524
The Lord of the Rings: The Two Towers	PG-13	179	Action	1346	908	1494

Which rating was the most popular among top grossing movies?

```r
ratingtop20gross <- top20gross %>%
  count(rating, sort = TRUE)
ggplot(ratingtop20gross, aes(rating, n)) +
  geom_col(fill = "blue") +
  labs(x = "Rating", y = "Number of Movies",
       title = "Which rating was the most popular among the top 20 grossing movies (actual)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```

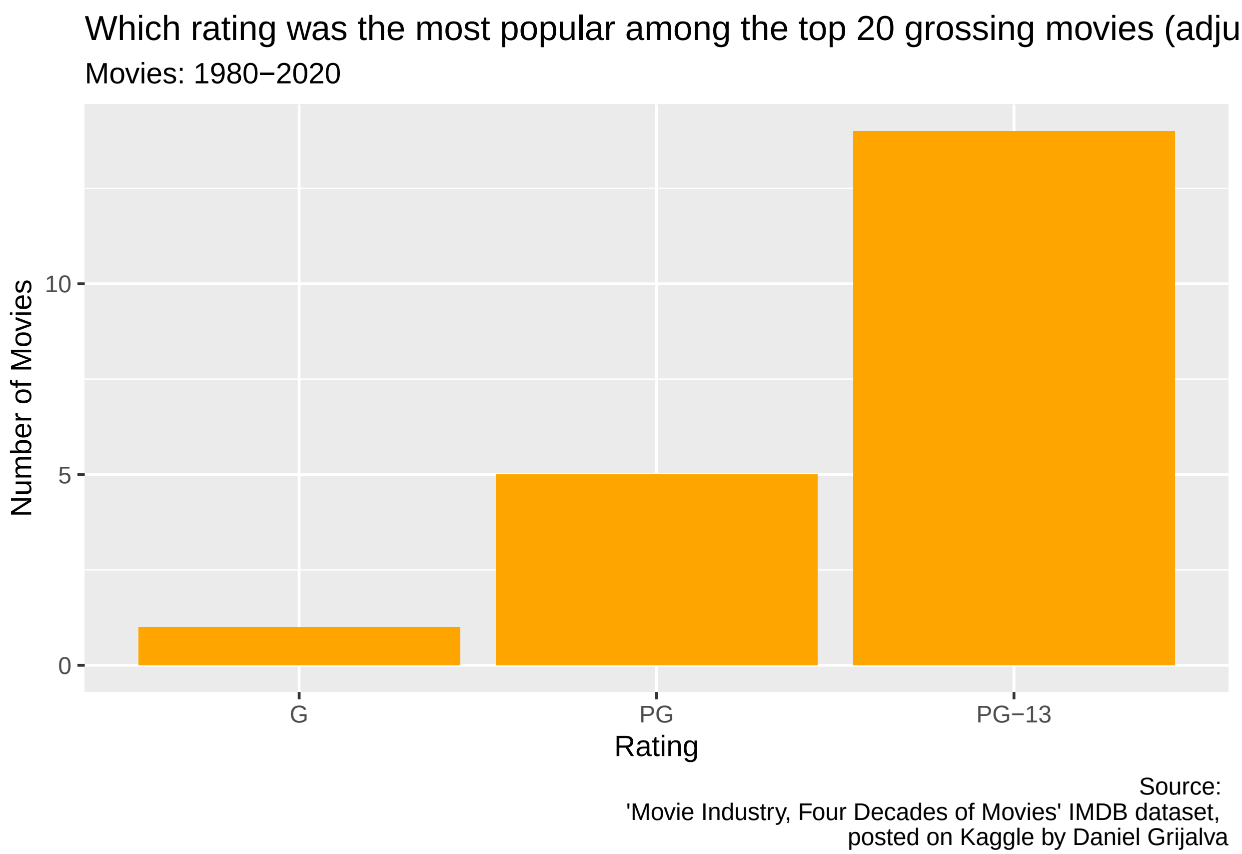


PG-13 movies dominated the Top 20 Highest Grossing Movie list with 15.

PG movies have 5, and R movies have 0.

Which rating was the most popular among top grossing movies (adjusted)?

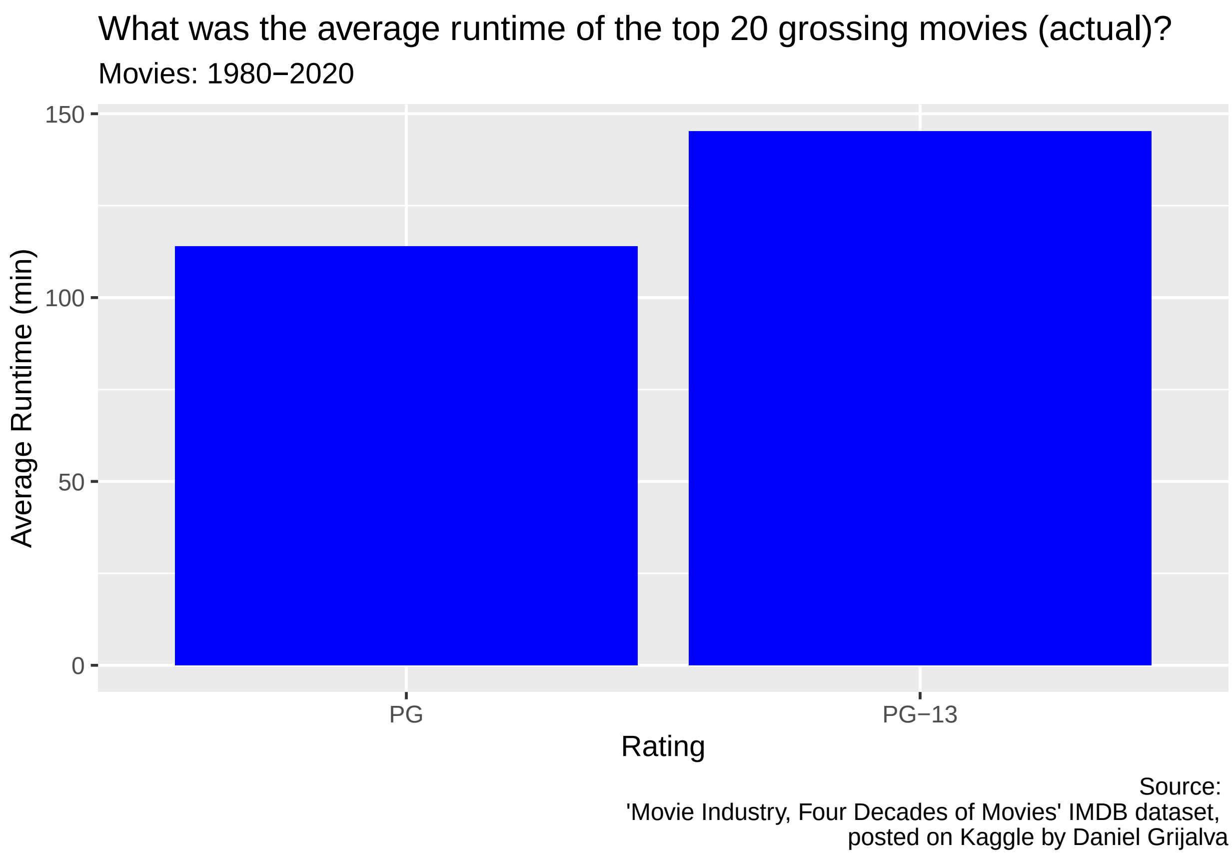
```r
adjratingtop20gross <- adjtop20gross %>%
  count(rating, sort = TRUE)
ggplot(adjratingtop20gross, aes(rating, n)) +
  geom_col(fill = "orange") +
  labs(x = "Rating", y = "Number of Movies",
       title = "Which rating was the most popular among the top 20 grossing movies (adjusted)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



There was one G rated movie in the top 20 grossing movies (adjusted), and still no R rated movies.

What was the average runtime among the top 20 grossing movies (actual)?

```
runtimetop20gross <- top20gross %>%
  group_by(rating) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
ggplot(runtimetop20gross, aes(rating, AVGruntime)) +
  geom_col(fill = "blue") +
  labs(x = "Rating", y = "Average Runtime (min)",
       title = "What was the average runtime of the top 20 grossing movies (actual)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```

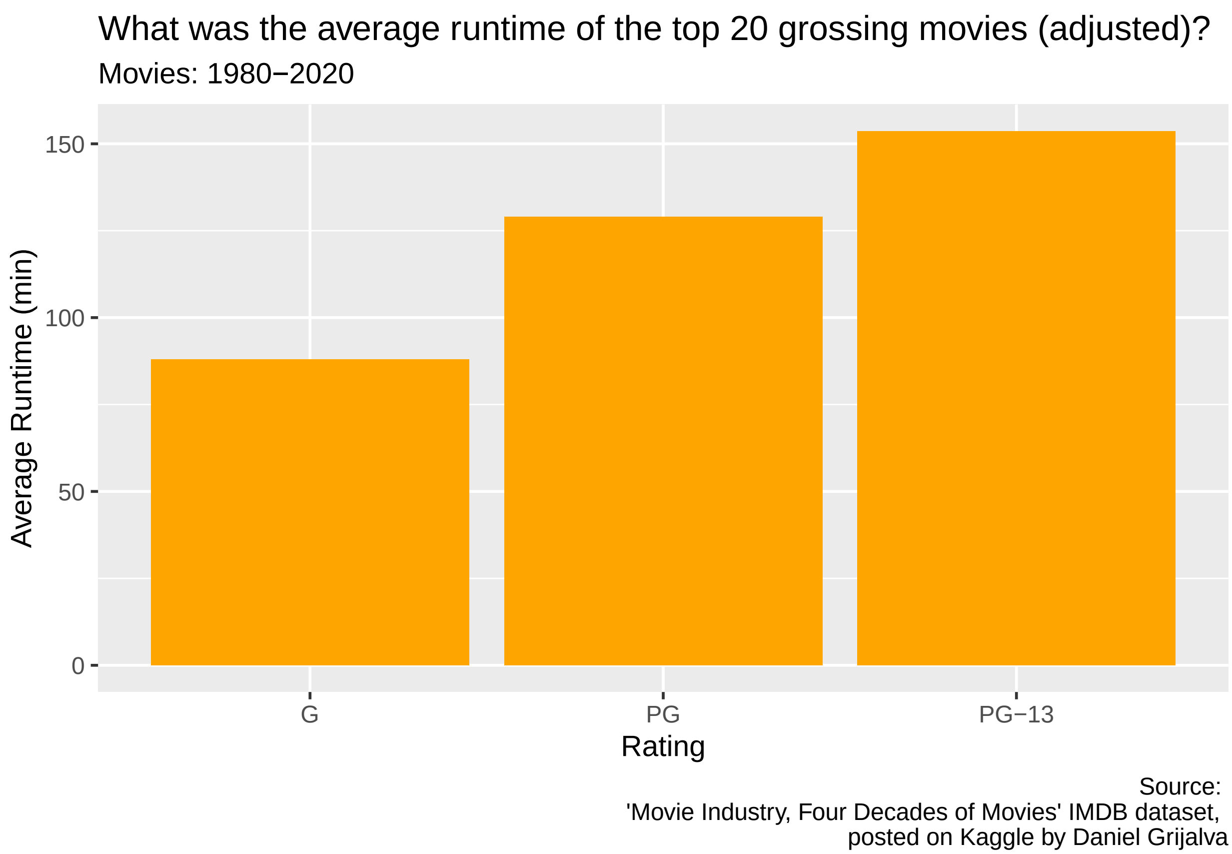


For PG-13, the avg runtime of the top 20 grossing movies was 145 min, which was 36 min longer than the avg of all PG-13 movies.

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 was the average runtime among the top 20 grossing movies (adjusted)?

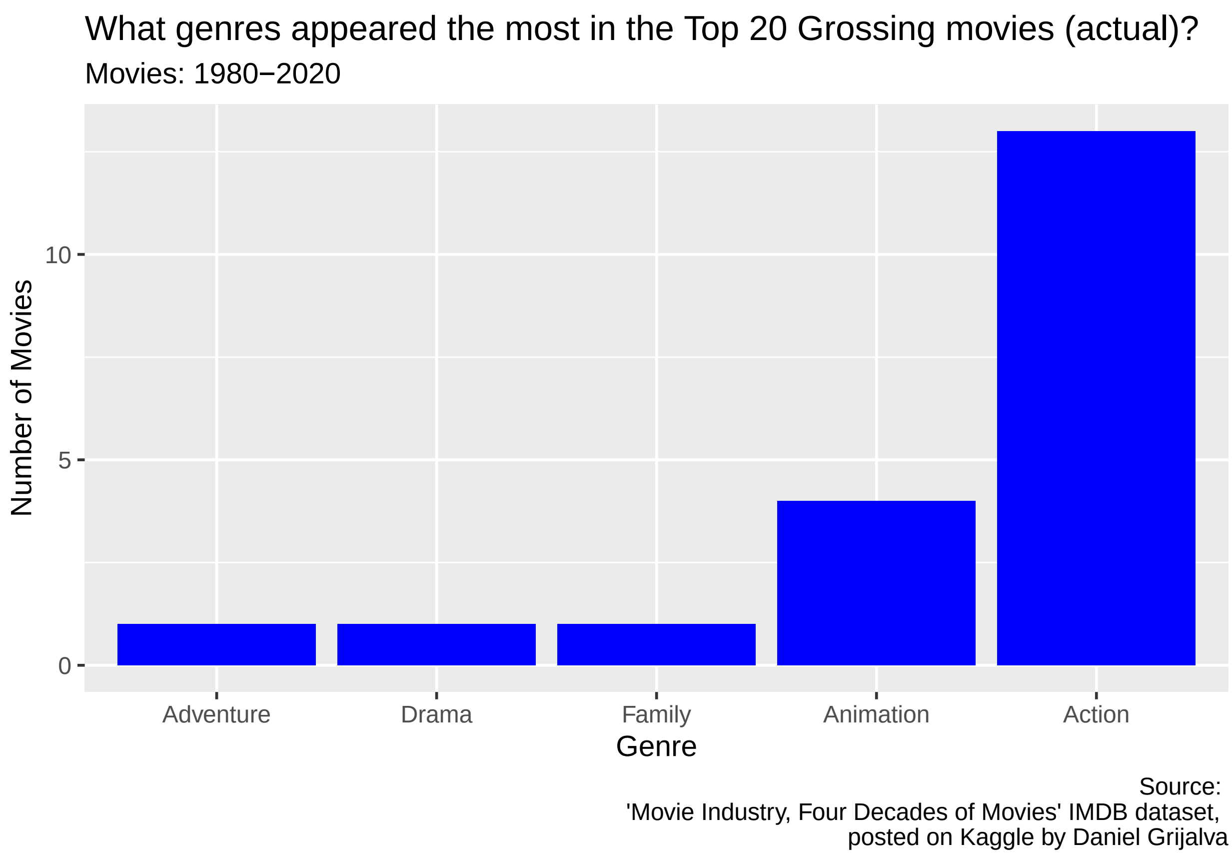
```
adjruntimetop20gross <- adjtop20gross %>%
  group_by(rating) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
ggplot(adjruntimetop20gross, aes(rating, AVGruntime)) +
  geom_col(fill = "orange") +
  labs(x = "Rating", y = "Average Runtime (min)",
       title = "What was the average runtime of the top 20 grossing movies (adjusted)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



The top 20 grossing movies (inflation-adjusted) ran longer than their "actual" counterparts.

What genres appeared the most in the Top 20 Grossing movies (actual)?

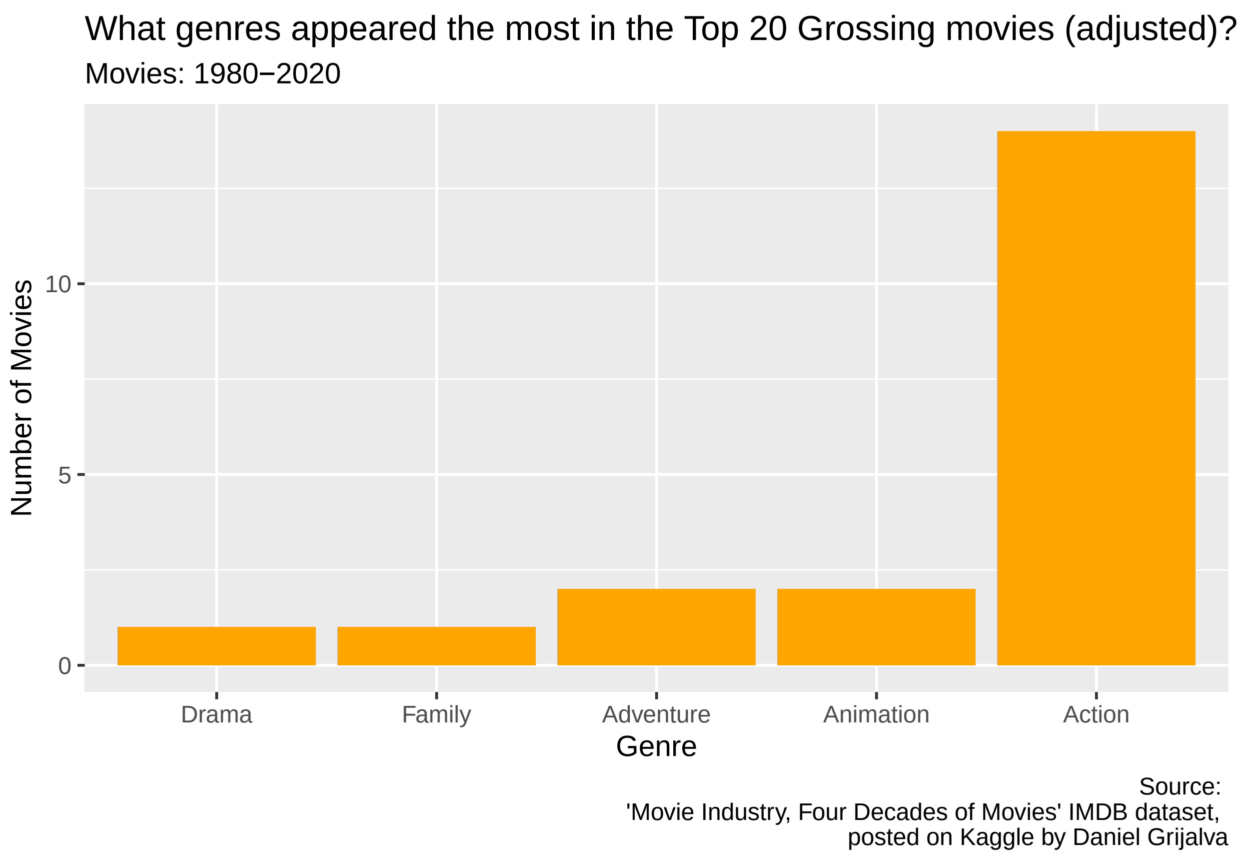
```
genretop20gross <- top20gross %>%
  group_by(genre) %>%
  count(genre, sort = TRUE)
ggplot(genretop20gross, aes(x = reorder(genre, n), y = n)) +
  geom_col(fill = "blue") +
  labs(x = "Genre", y = "Number of Movies",
       title = "What genres appeared the most in the Top 20 Grossing movies (actual)?",
       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 (actual).

What genres appeared the most in the Top 20 Grossing movies (adjusted)?

```
adjgenretop20gross <- adjtop20gross %>%
  group_by(genre) %>%
  count(genre, sort = TRUE)
ggplot(adjgenretop20gross, aes(x = reorder(genre, n), y = n)) +
  geom_col(fill = "orange") +
  labs(x = "Genre", y = "Number of Movies",
       title = "What genres appeared the most in the Top 20 Grossing movies (adjusted)?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```



The genre distribution was similar between actual and adjusted top 20 grossing movies, with slightly less Animation and slightly more Action for the adjusted movies.

Top Profitable Analysis of Select Variables

Top 20 Movies with Highest Profit Percentage:

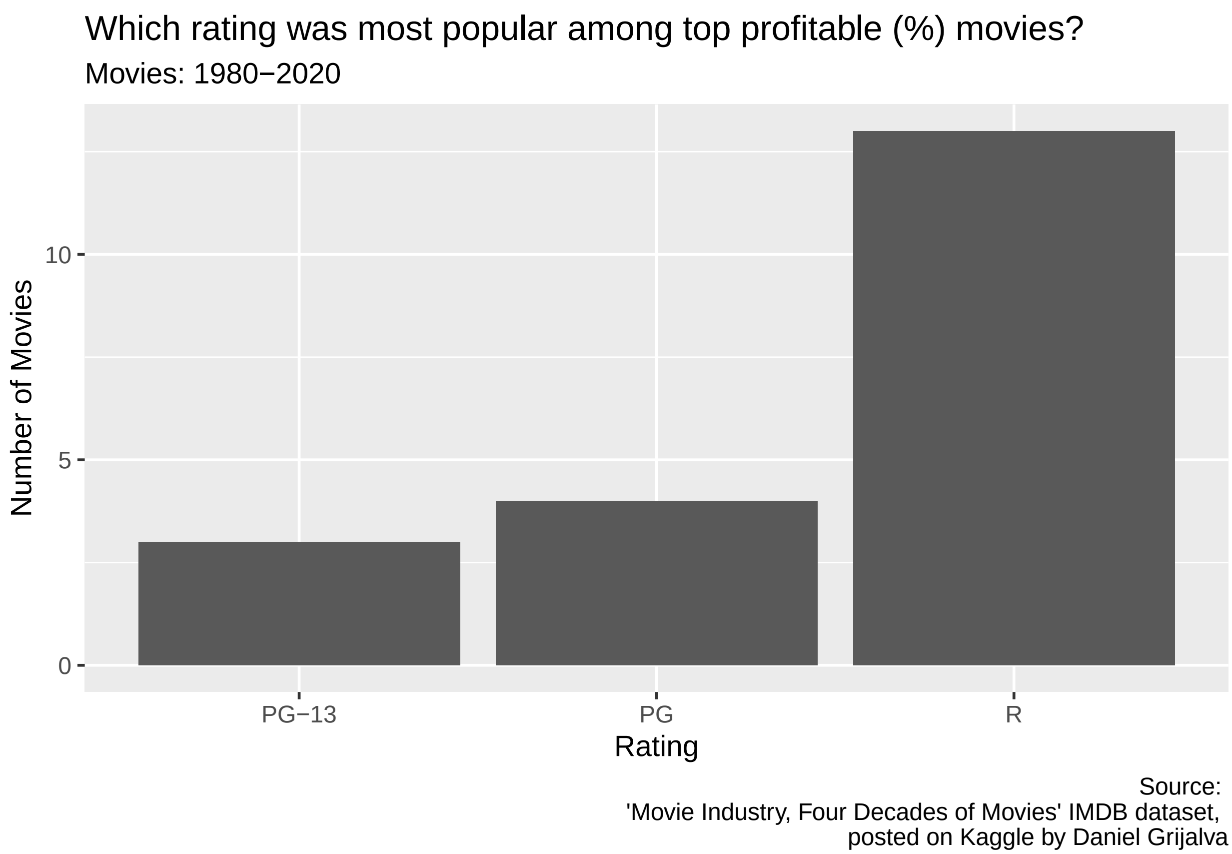
```
top20profperc <- movies %>%
  select(c(name, rating, runtime, genre, profitM, adjprofitM, profit_percent)) %>%
  arrange(desc(profit_percent)) %>%
  top_n(20)
top20profperc %>%
  knitr::kable(caption = "Top 20 Profitable Movies: Categorical", digits = 0) %>%
  kableExtra::kable_styling(latex_options = c("hold_position"))
```

Table 10: Top 20 Profitable Movies: Categorical

name	rating	runtime	genre	profitM	adjprofitM	profit_percent
Paranormal Activity	R	86	Horror	193	257	1288939
The Blair Witch Project	R	81	Horror	249	448	414298
The Gallows	R	81	Horror	43	47	42864
El Mariachi	R	81	Action	2	4	29056
Once	R	86	Drama	21	28	13858
Clerks	R	92	Comedy	3	7	11571
Napoleon Dynamite	PG	96	Comedy	46	67	11435
In the Company of Men	R	97	Comedy	3	6	11118
Keeping Mum	R	99	Comedy	18	26	10898
Open Water	R	79	Adventure	54	82	10837
The Devil Inside	R	83	Horror	101	116	10076
The Quiet Ones	PG-13	98	Horror	18	20	8818
Saw	R	103	Horror	103	152	8559
Searching	PG-13	102	Drama	75	75	8475
Primer	PG-13	77	Drama	1	1	7692
E.T. the Extra-Terrestrial	PG	115	Family	782	2438	7452
My Big Fat Greek Wedding	PG	95	Comedy	364	573	7275
The Full Monty	R	91	Comedy	254	508	7270
Friday the 13th	R	95	Horror	39	133	7128
Fireproof	PG	122	Drama	33	42	6595

Which rating was most popular among top profitable (%) movies?

```
ratingtop20profperc <- top20profperc %>%
  count(rating, sort = TRUE)
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 profitable (%) movies?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```

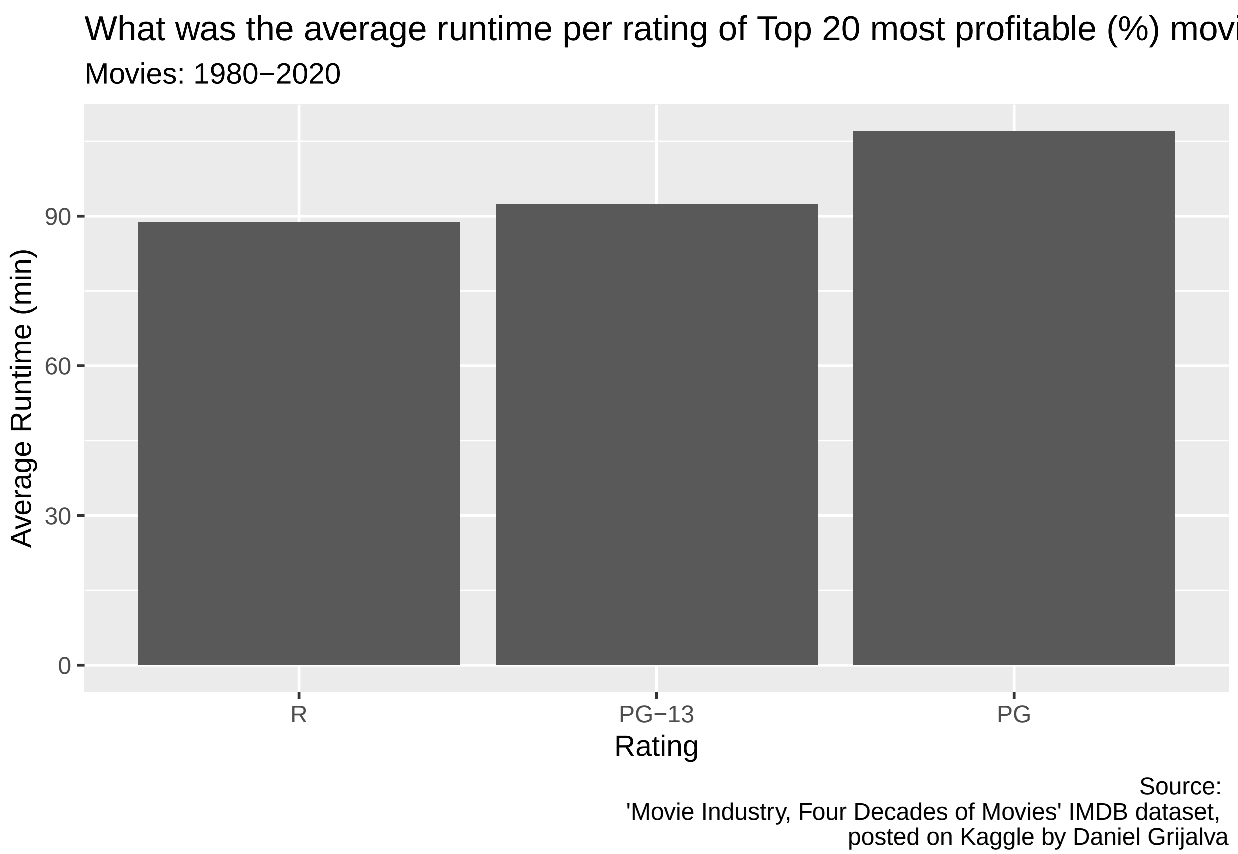


R movies dominated the top profit (%) movie list with 13.

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

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

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

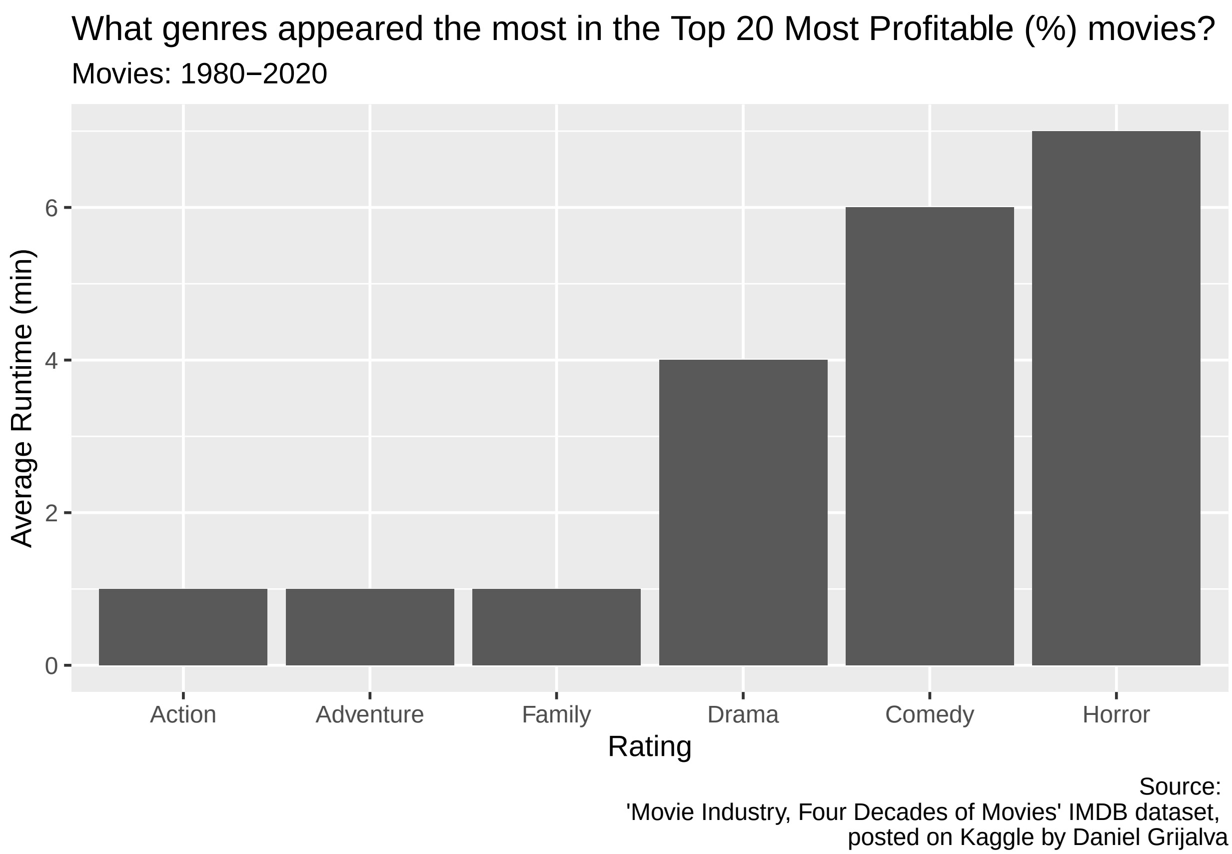
For R, the avg runtime of the top 20 most profitable (%) movies was 89 min, which was 19 min less than the avg of all R movies.

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?

```
genretop20profperc<- top20profperc %>%
  group_by(genre) %>%
  count(genre, sort = TRUE)
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?",
       subtitle = "Movies: 1980-2020",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel Grijalva")
```

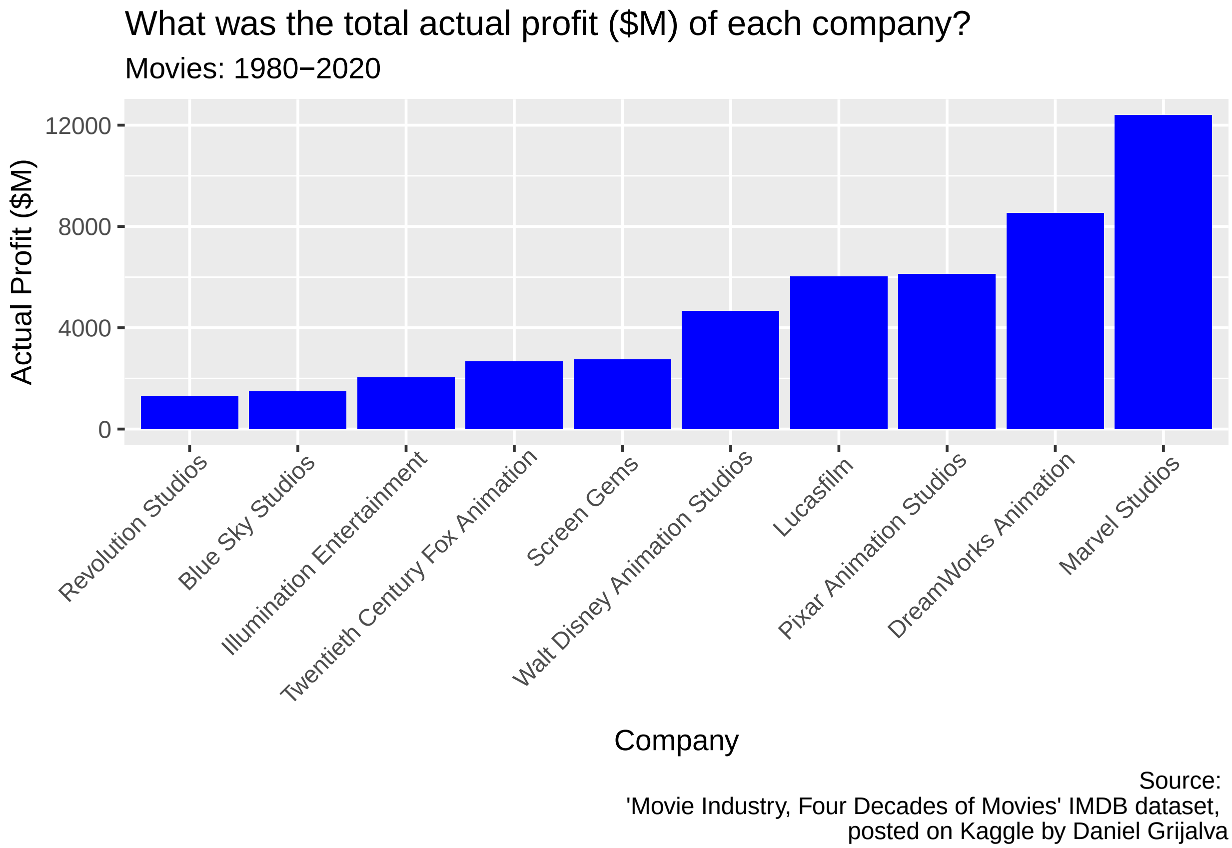


Horror had the most (7), followed by Comedy (6) and Drama (4).

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

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

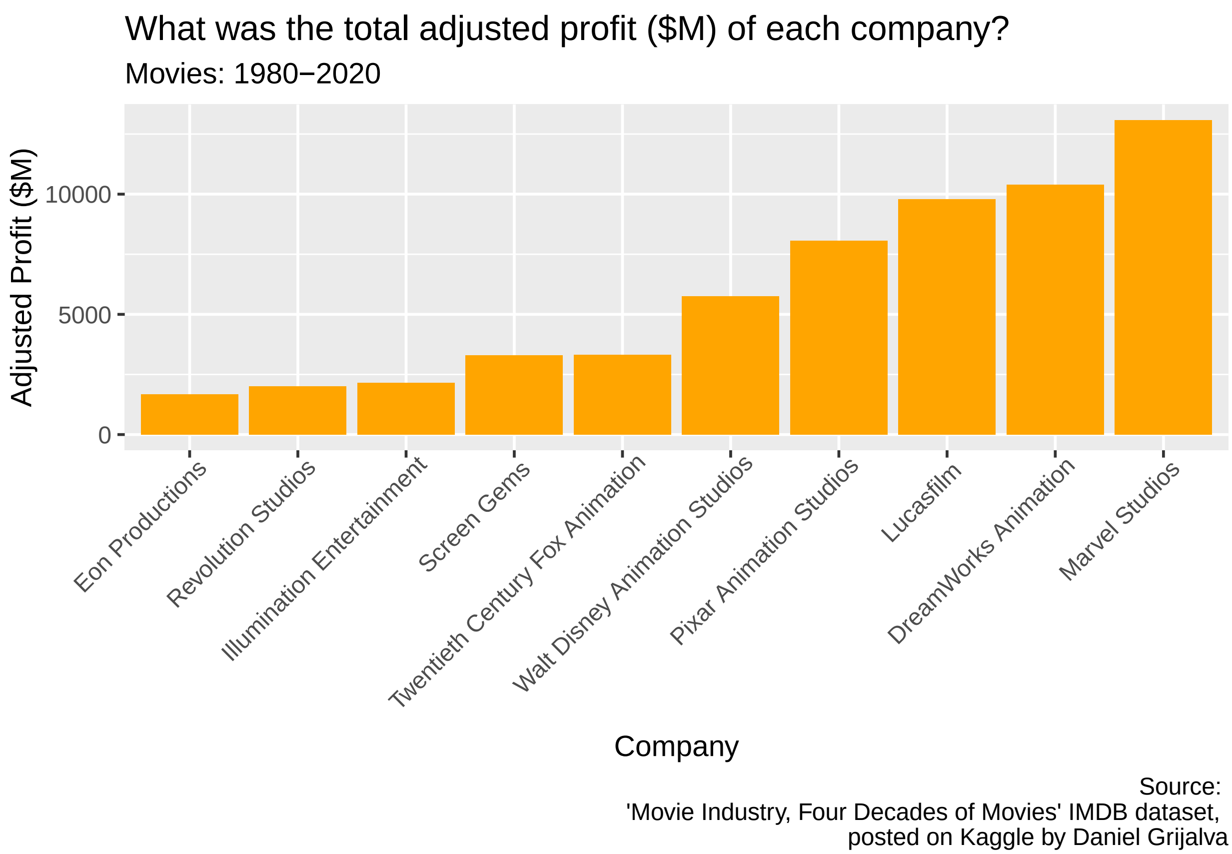
```
top10profitcomp <- movies %>%
  group_by(company) %>%
  summarise(profitM = sum(profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(10)
ggplot(top10profitcomp, aes(x = reorder(company, profitM), y = profitM)) +
  geom_col(fill = "blue") +
  labs(x = "Company", y = "Actual Profit ($M)",
       title = "What was the total actual profit ($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 actual profit by far ($12B), followed by Dreamworks ($8B), Pixar ($6B), and Lucasfilm ($6B).

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

```
adjtop10profitcomp <- movies %>%
  group_by(company) %>%
  summarise(adjprofitM = sum(adjprofitM)) %>%
  arrange(desc(adjprofitM)) %>%
  top_n(10)
ggplot(adjtop10profitcomp, aes(x = reorder(company, adjprofitM), y = adjprofitM)) +
  geom_col(fill = "orange") +
  labs(x = "Company", y = "Adjusted Profit ($M)",
       title = "What was the total adjusted profit ($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))
```

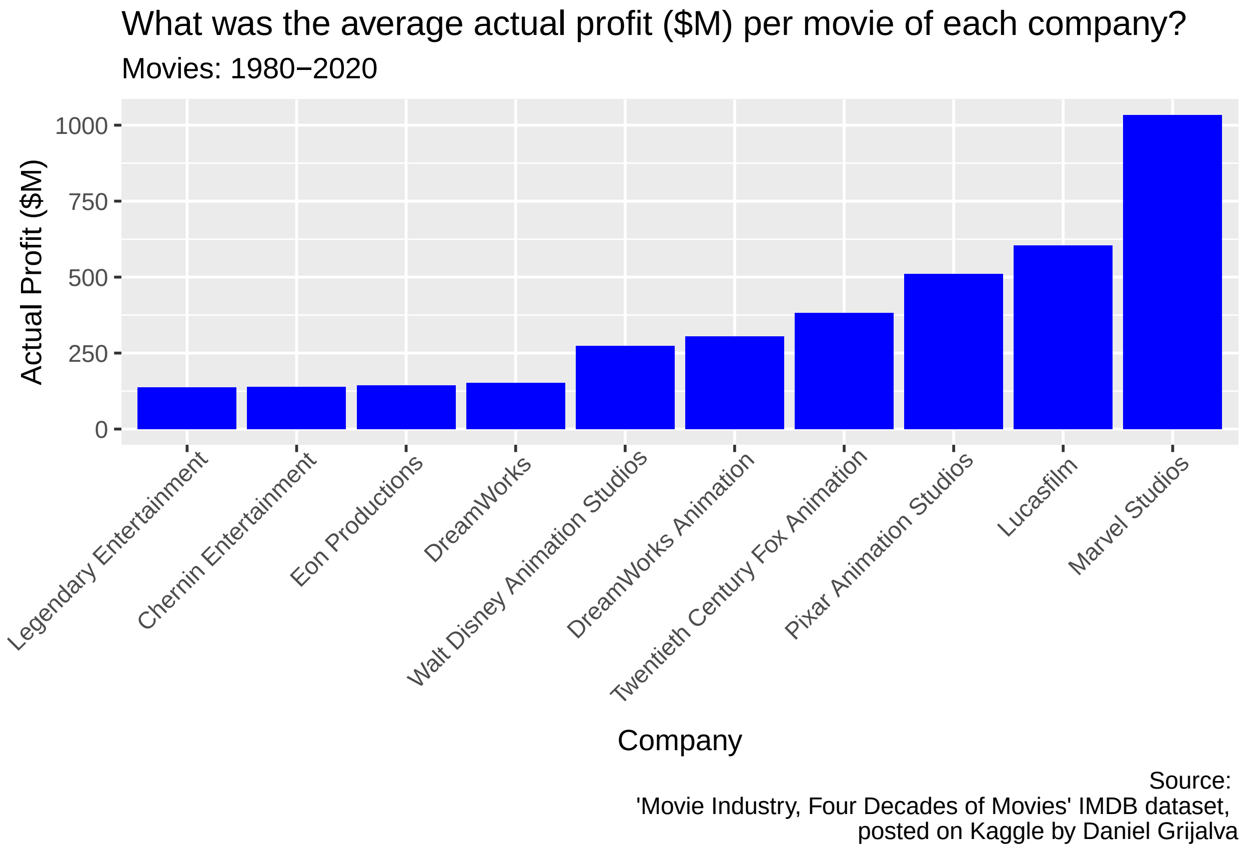


Marvel tops both the actual and adjusted profit list, but both Dreamworks and Lucasfilm have higher adjusted profits than their acutal profits.

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

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

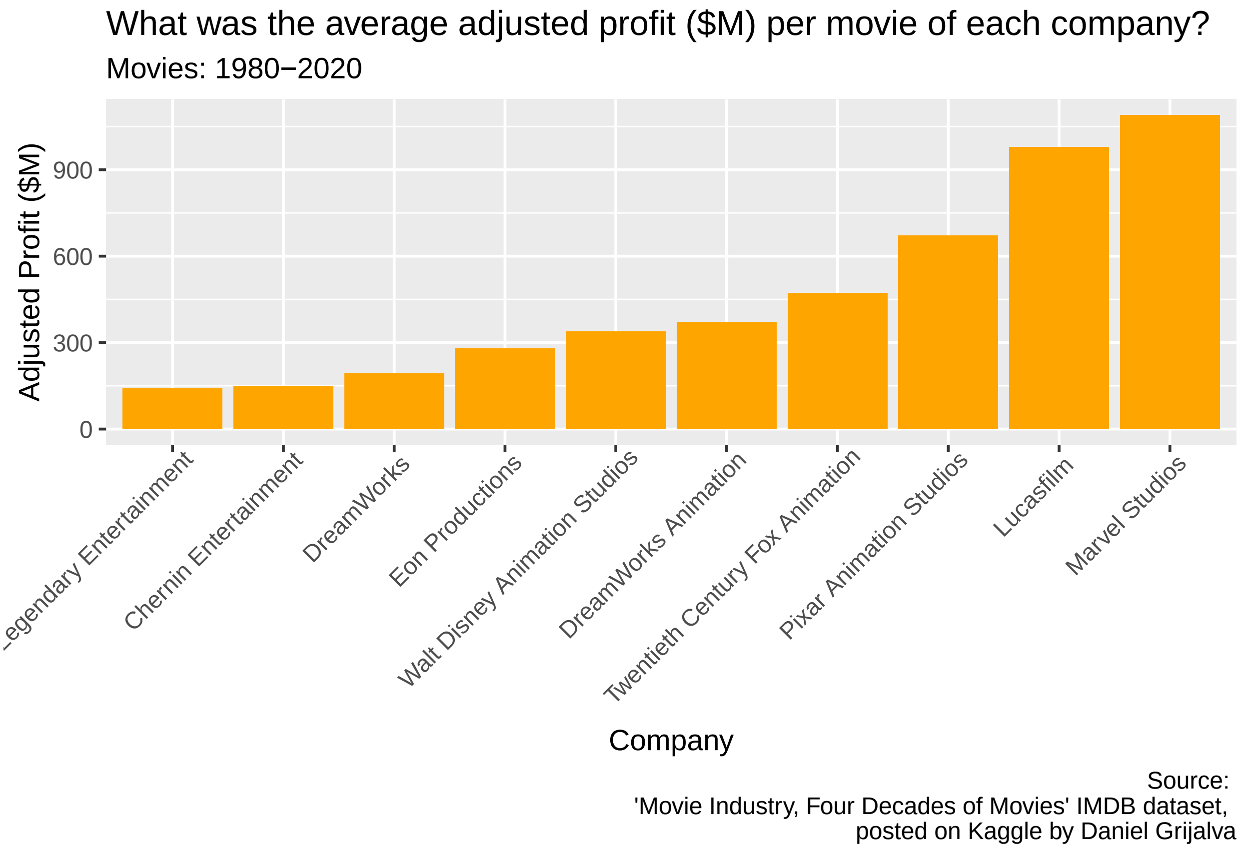
```
compN <- movies %>%
  group_by(company) %>% filter(n() >= 5) %>% ungroup() #include companies with >= 5 movies)
profcompN <- compN %>%
  group_by(company) %>%
  summarise(profitM = mean(profitM)) %>%
  arrange(desc(profitM)) %>%
  top_n(10)
ggplot(profcompN, aes(x = reorder(company, profitM), y = profitM)) +
  geom_col(fill = "blue") +
  labs(x = "Company", y = "Actual Profit ($M)",
       title = "What was the average actual profit ($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 adjusted profit ($M) per movie of each company?

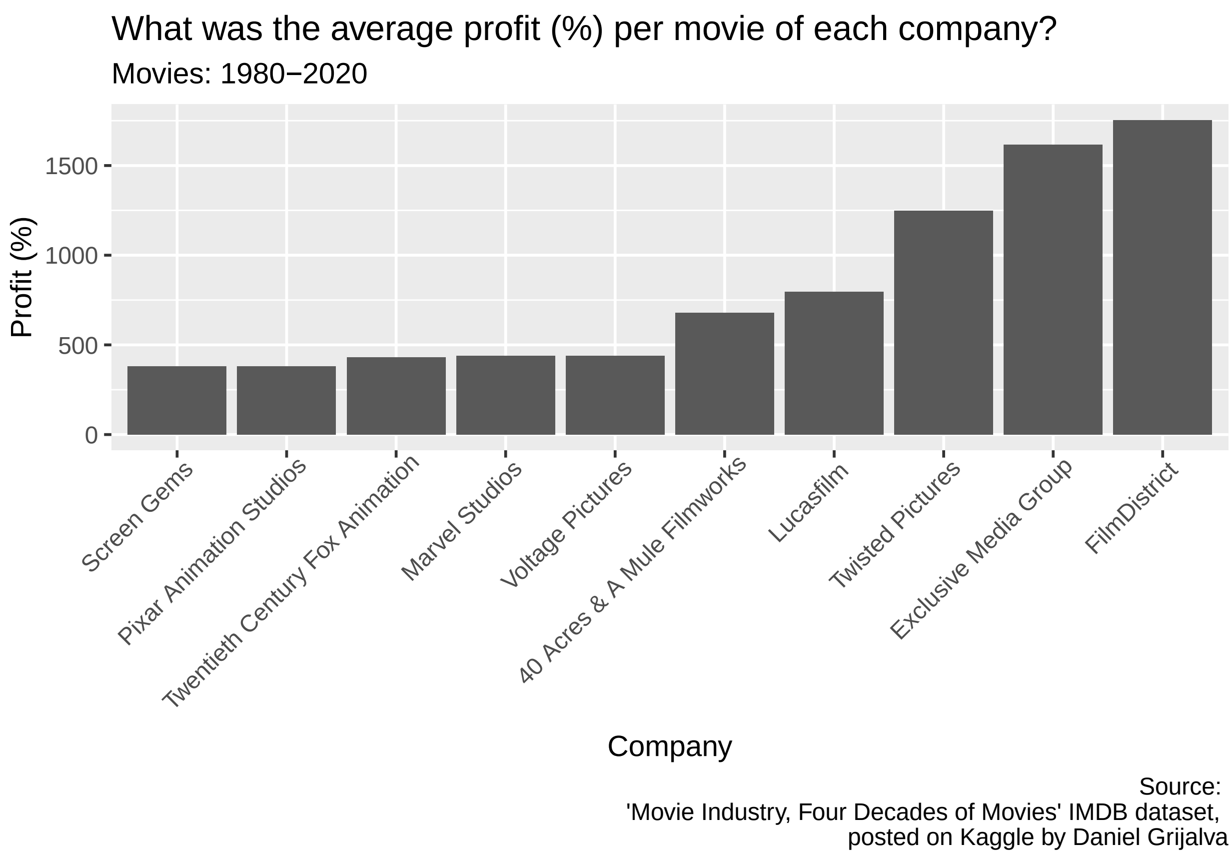
```
adjprofcompN <- compN %>%
  group_by(company) %>%
  summarise(adjprofitM = mean(adjprofitM)) %>%
  arrange(desc(adjprofitM)) %>%
  top_n(10)
ggplot(adjprofcompN, aes(x = reorder(company, adjprofitM), y = adjprofitM)) +
  geom_col(fill = "orange") +
  labs(x = "Company", y = "Adjusted Profit ($M)",
       title = "What was the average adjusted profit ($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 still tops the average adjusted profit ($M) list, however the gap between it and Lucasfilm is much narrower compared to the actual list.

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)
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)?

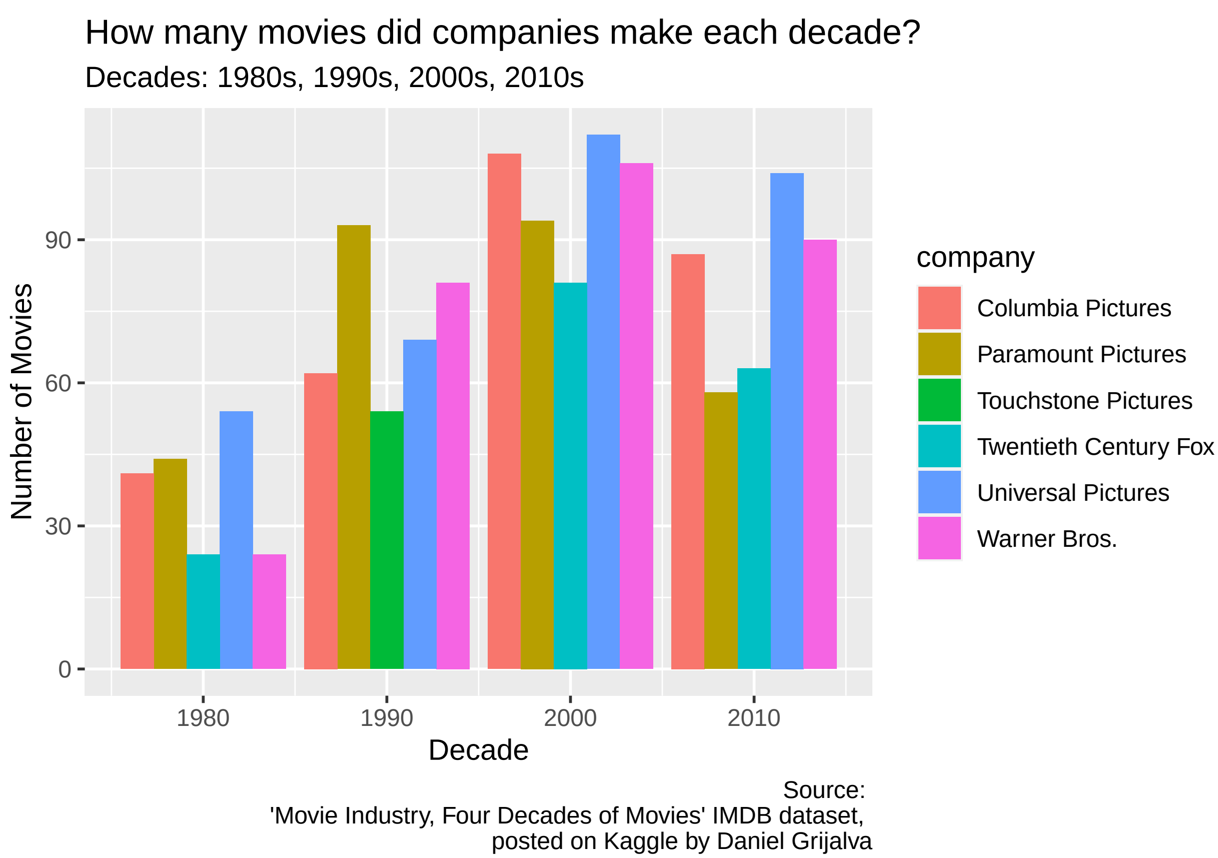
```
compDecade <- movies %>%
  filter(!is.na(decade), decade != 2020) %>% # eliminate NA decades and 2020 films
  group_by(decade, company) %>%
  count(decade, sort = TRUE)
compDecade <- compDecade %>%
  arrange(desc(n)) %>%
  group_by(decade) %>%
  slice(1:5)  # Top 5 highest values (number of movies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

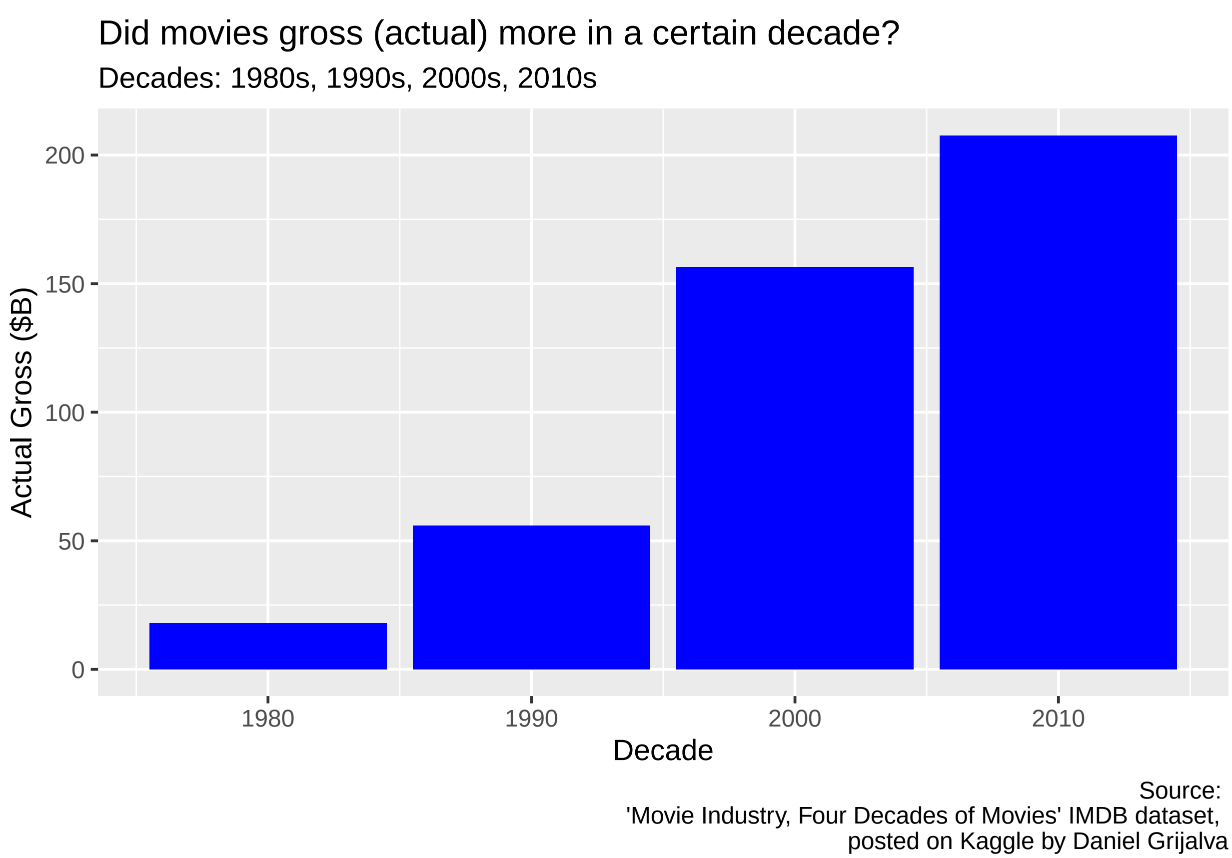
Plot via clustered column chart:

```
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 (actual) more in a certain decade?

```
grossdecade <- movies %>%
  filter(!is.na(grossM), decade != 2020) %>%
  group_by(decade) %>%
  summarise(grossB = sum(grossM) / 1000) %>%
  arrange(desc(grossB))
ggplot(grossdecade, aes(x = decade, y = grossB)) +
  geom_col(fill = "blue") +
  labs(x = "Decade", y = "Actual Gross ($B)",
       title = "Did movies gross (actual) 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")
```

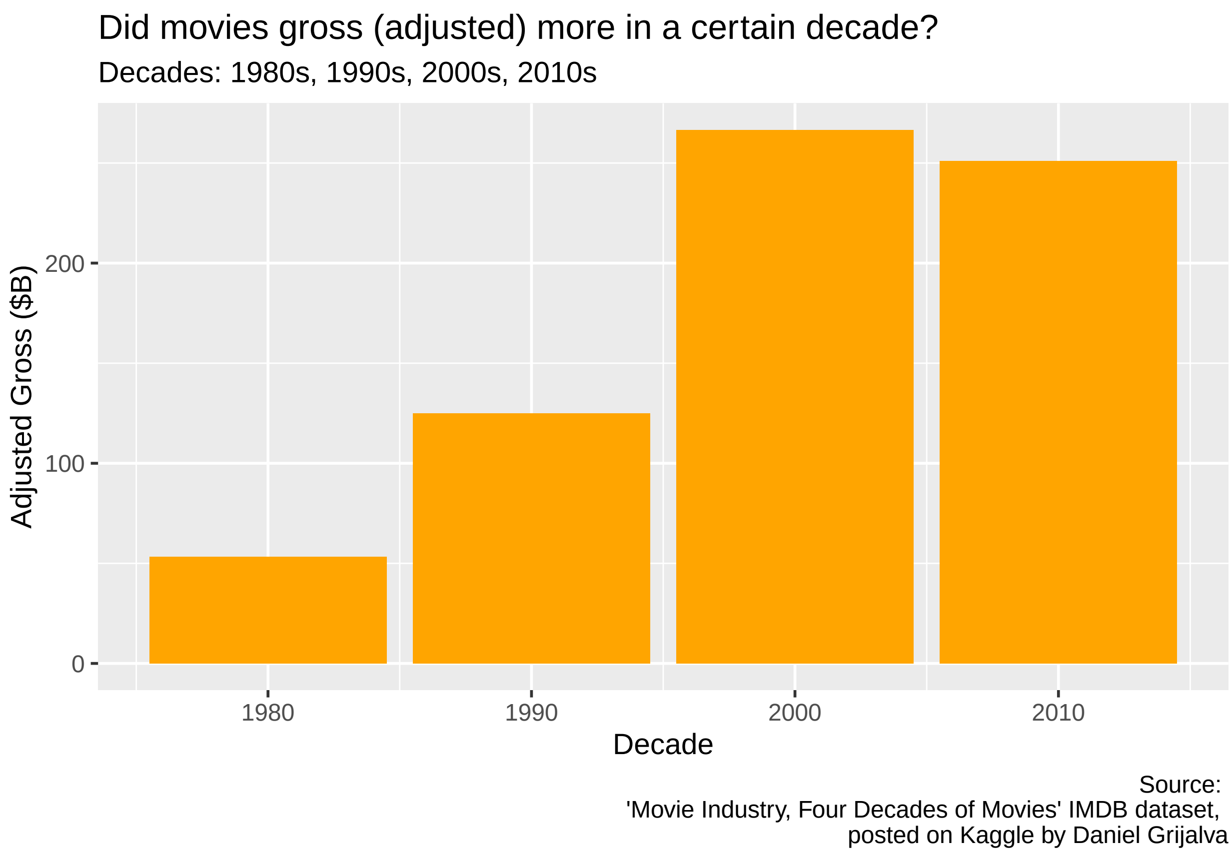


There was a cumulative rise in gross each decade. The cumulative rise in movie ticket prices is one possible factor.

Did movies gross (adjusted) more in a certain decade?

```
adjgrossdecade <- movies %>%
  filter(!is.na(adjgrossM), decade != 2020) %>%
  group_by(decade) %>%
  summarise(adjgrossB = sum(adjgrossM) / 1000) %>%
  arrange(desc(adjgrossB))
ggplot(adjgrossdecade, aes(x = decade, y = adjgrossB)) +
  geom_col(fill = "orange") +
```

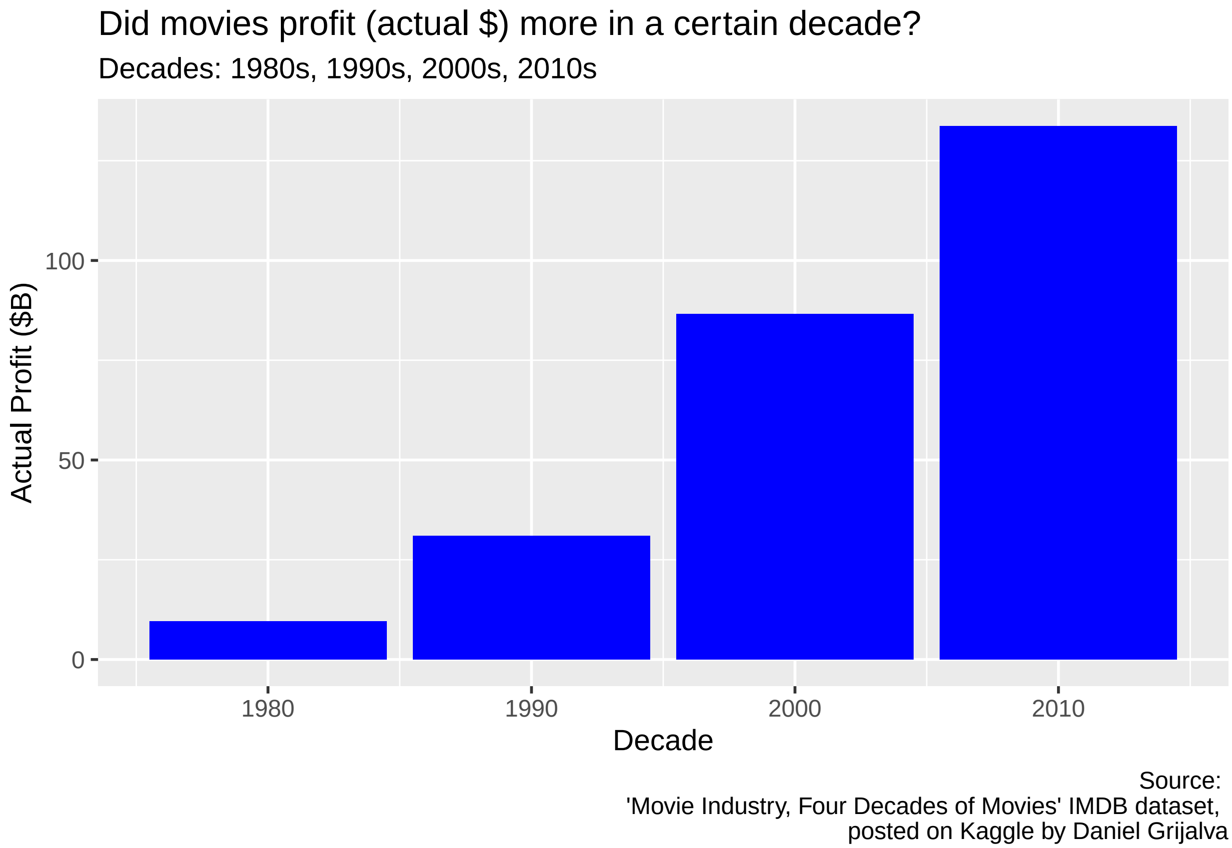
```r
  labs(x = "Decade", y = "Adjusted Gross ($B)",
       title = "Did movies gross (adjusted) more in a certain decade?",
       subtitle = "Decades: 1980s, 1990s, 2000s, 2010s",
       caption = "Source:
       'Movie Industry, Four Decades of Movies' IMDB dataset,
       posted on Kaggle by Daniel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



When considering inflation, movies grossed more in the 2000s than any other decade.

Did movies profit (actual $) more in a certain decade?

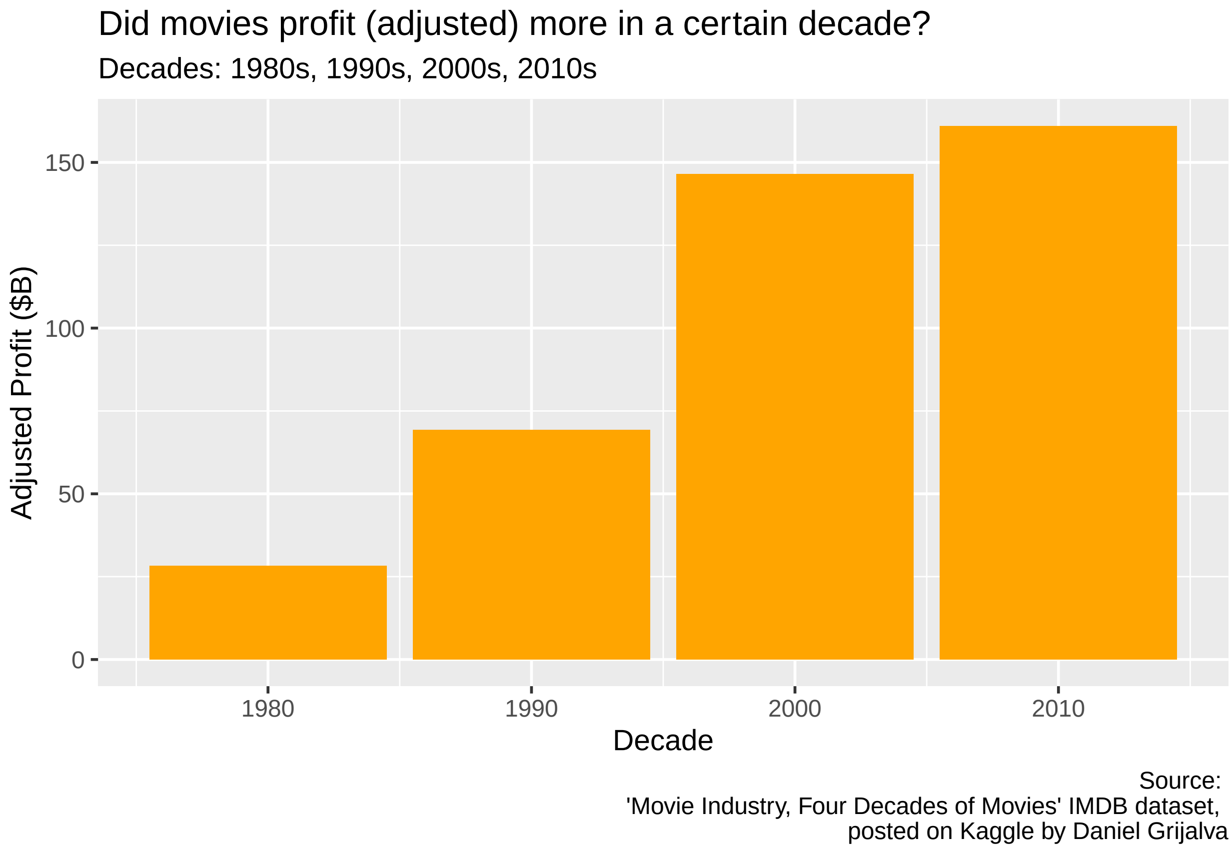
```r
profdecade <- movies %>%
  filter(!is.na(profitM), decade != 2020) %>%
  group_by(decade) %>%
  summarise(profitB = sum(profitM) / 1000) %>%
  arrange(desc(profitB))
ggplot(profdecade, aes(x = decade, y = profitB)) +
  geom_col(fill = "blue") +
  labs(x = "Decade", y = "Actual Profit ($B)",
       title = "Did movies profit (actual $) 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 actual profit ($) each decade. The cumulative rise in movie ticket prices is one possible factor.

Did movies profit (adjusted $) more in a certain decade?

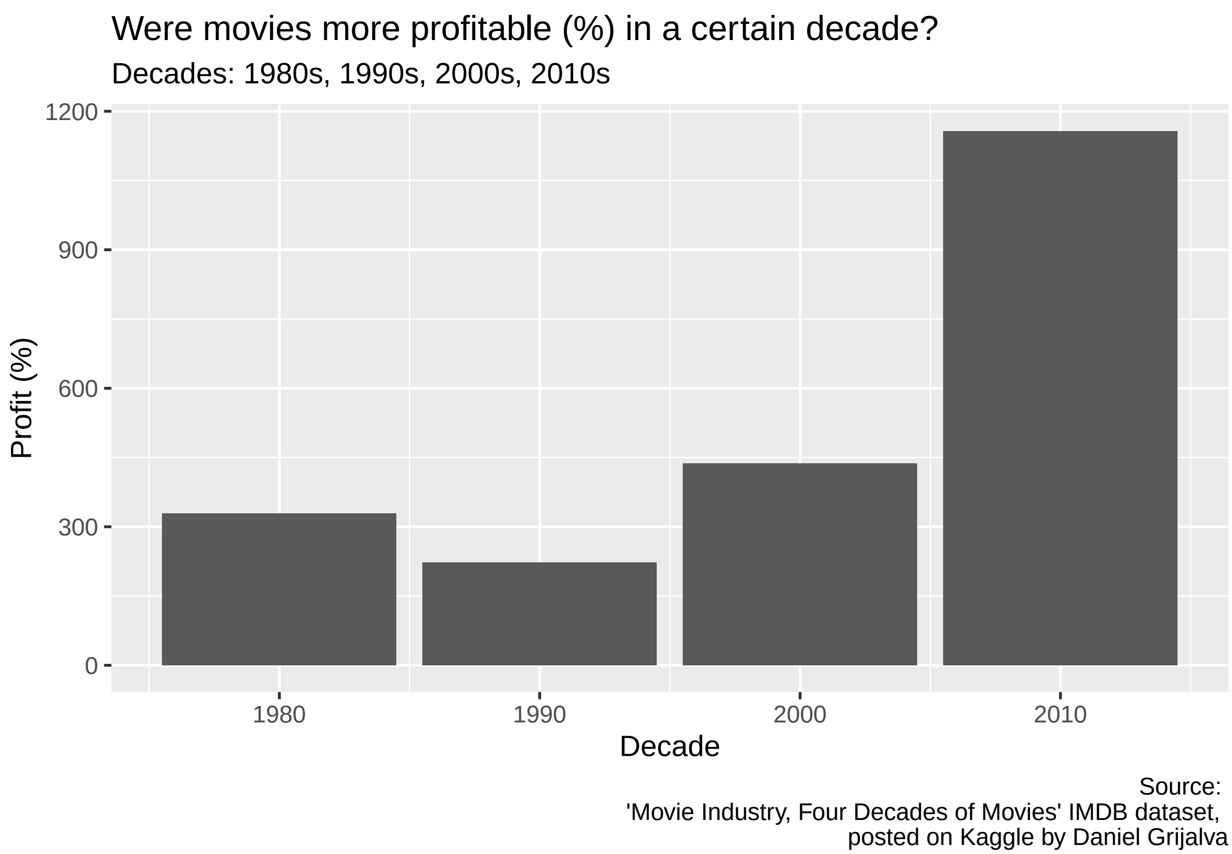
```
adjprofdecade <- movies %>%
  filter(!is.na(adjprofitM), decade != 2020) %>%
  group_by(decade) %>%
  summarise(adjprofitB = sum(adjprofitM) / 1000) %>%
  arrange(desc(adjprofitB))
ggplot(adjprofdecade, aes(x = decade, y = adjprofitB)) +
  geom_col(fill = "orange") +
  labs(x = "Decade", y = "Adjusted Profit ($B)",
       title = "Did movies profit (adjusted) 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")
```



When considering inflation, movies still profited ($) more in the 2010s, however the gap between 2010s and 2000s is much narrower compared to the actual profit list.

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

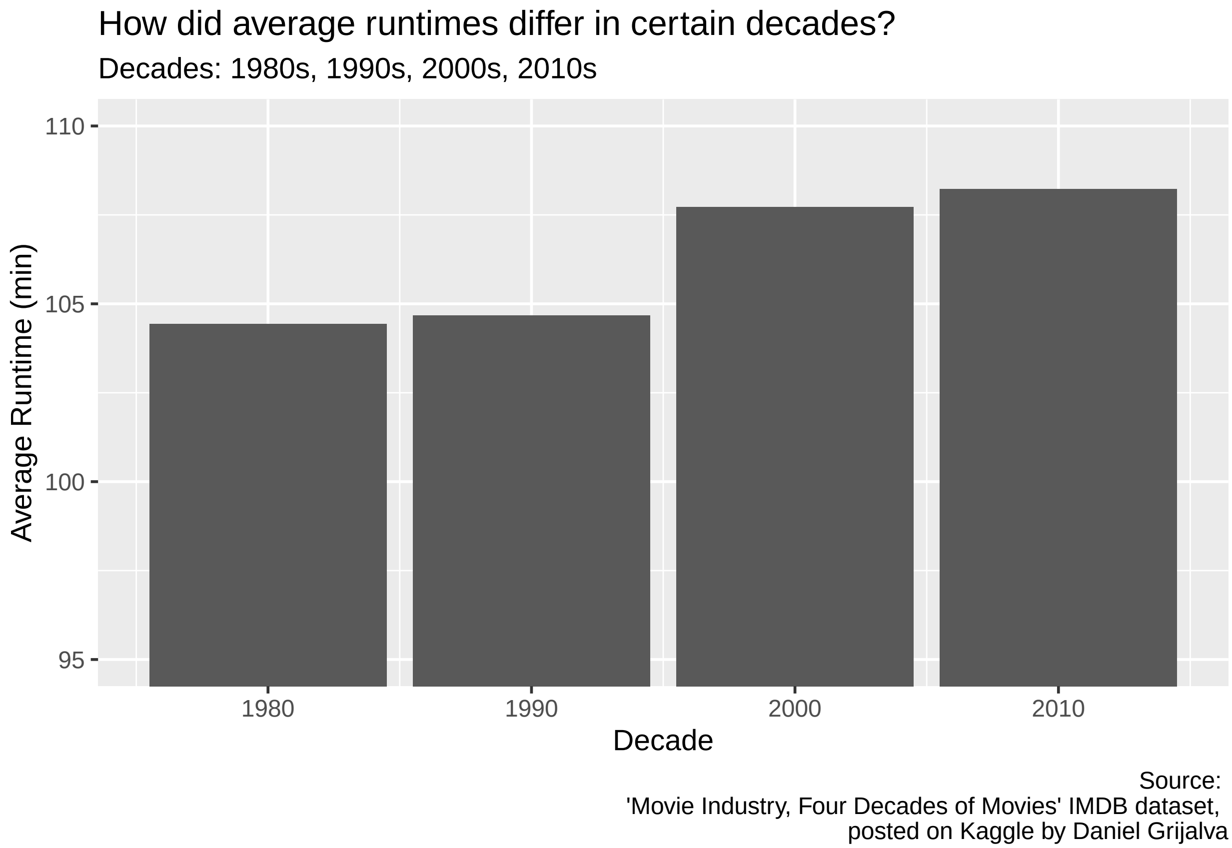
```
percdecade <- movies %>%
  filter(!is.na(profit_percent), decade != 2020) %>%
  group_by(decade) %>%
  summarise(profit_percent = mean(profit_percent)) %>%
  arrange(desc(profit_percent))
ggplot(percdecade, aes(x = decade, y = profit_percent)) +
  geom_col() +
  labs(x = "Decade", y = "Profit (%)",
       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?

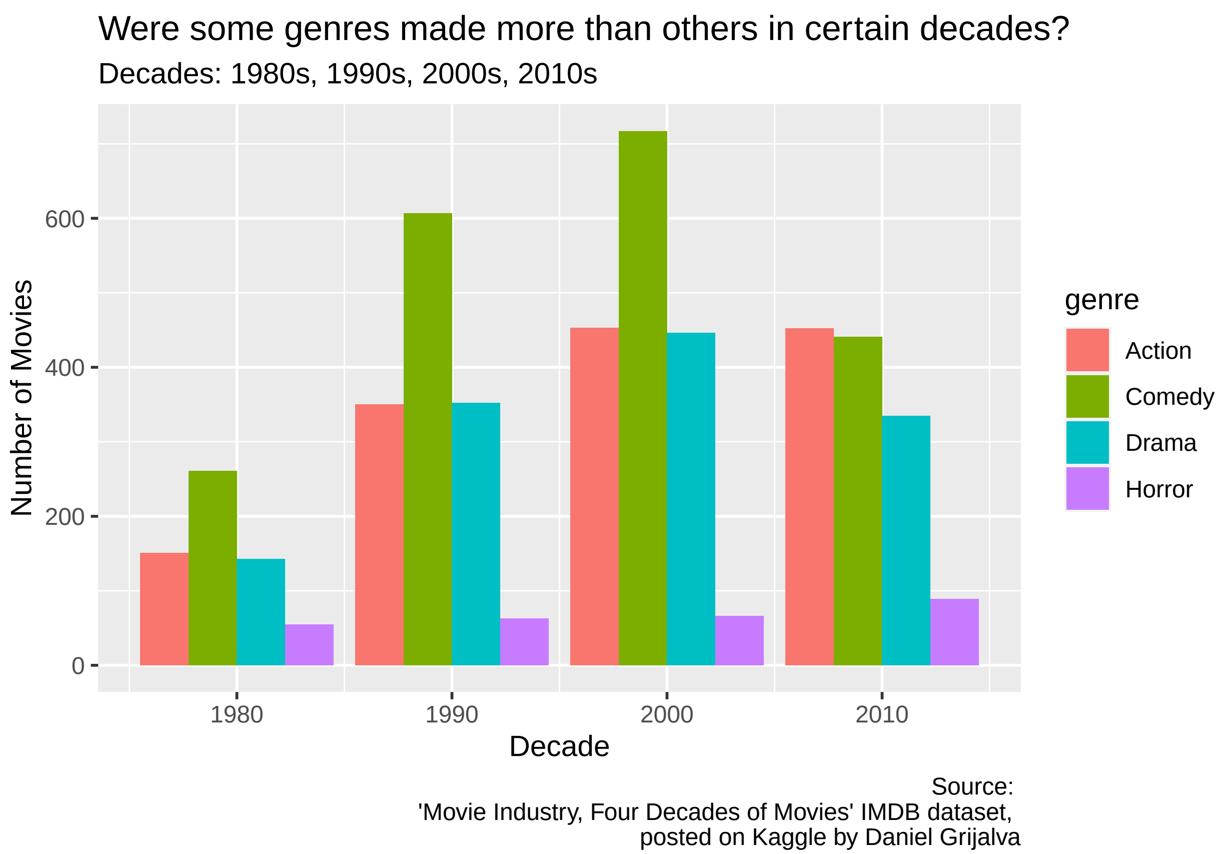
```
runtimedecade <- movies %>%
  filter(!is.na(runtime), decade != 2020) %>%
  group_by(decade) %>%
  summarise(AVGruntime = mean(runtime)) %>%
  arrange(desc(AVGruntime))
ggplot(runtimedecade, aes(x = decade, y = AVGruntime)) +
  geom_col() +
  labs(x = "Decade", y = "Average Runtime (min)",
       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)
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.

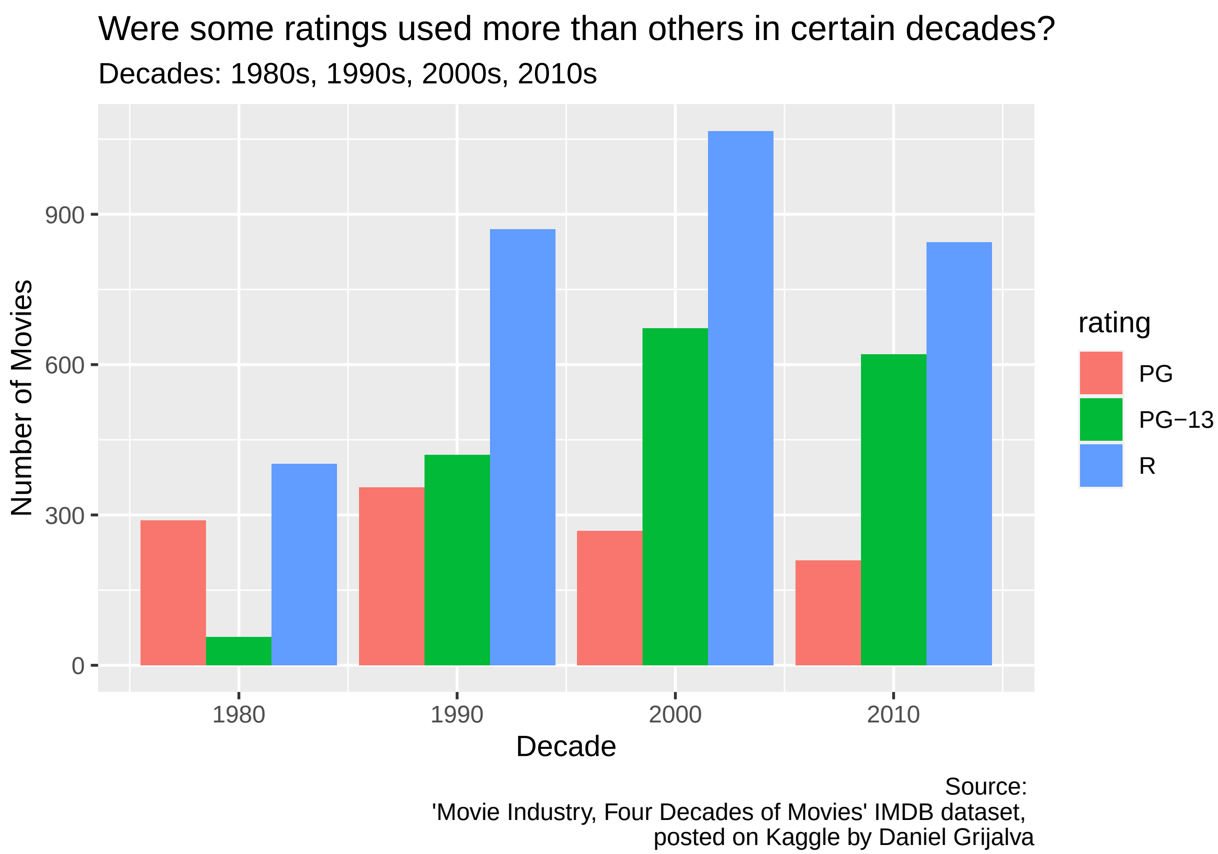
More comedy movies were made in 90s-00s than in 80s/10s.

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)
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.
- Adjusting for inflation, the top 3 grossing movies were mostly the same (Titanic, Avatar, Avengers:Endgame). However E.T., The Lion King (1994), and Jurassic Park were highly successful for their times, and none were on the Top 20 Actual Grossing list. None of the top 6 inflation-adjusted grossing movies were franchise related at the time of their release except Avengers:Endgame. The top 20 adjusted grossing movies were more spread out over the decades than the top 20 actual grossing movies.
- The Top 20 profitable (%) movies were even 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 actual grossing films; Horror and Comedy combined for 13 out of the top 20 profitable (%) films.
- Marvel had the most actual profit of all film companies ($12B), followed by Dreamworks ($8B), Pixar ($6B), and Lucasfilm ($6B). Marvel made about $1B profit per movie.
- When considering for inflation, Dreamworks and Lucasfilm have significantly higher profits but still do not top Marvel.
- Marvel averages the most per movie, but when considering for inflation, Lucasfilm is a much closer 2nd place.
- FilmDistrict, Exclusive Media Group, and Twisted Pictures all averaged over 1000% profit from their movies.
- There was a cumulative rise in actual gross among all movies each decade, but when considering for inflation the 2000s grossed the most.
- 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 %>%
  select(-c(adjgrossM, adjbudgetM, adjprofitM, inflation, AvgUSTicketPrice))
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):