DAT301 Lab2

2024-01-28

Load Data

```
library(ggplot2movies)
data(movies)
```

Load Packages

```
library(ggplot2)
library(dplyr)
```

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

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

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

Load Data set

```
head(movies)
```

```
## # A tibble: 6 x 24
##   title      year length budget rating votes    r1    r2    r3    r4    r5    r6
##   <chr>     <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 $          1971    121     NA    6.4   348   4.5   4.5   4.5   4.5  14.5  24.5
## 2 $1000 a ~  1939     71     NA    6      20   0    14.5   4.5  24.5  14.5  14.5
## 3 $21 a Da~  1941      7     NA    8.2     5   0     0     0     0     0    24.5
## 4 $40,000    1996     70     NA    8.2     6  14.5   0     0     0     0     0
## 5 $50,000 ~  1975     71     NA    3.4    17  24.5   4.5   0    14.5  14.5   4.5
## 6 $pent      2000     91     NA    4.3    45   4.5   4.5   4.5  14.5  14.5  14.5
## # ... with 12 more variables: r7 <dbl>, r8 <dbl>, r9 <dbl>, r10 <dbl>,
## #   mpaa <chr>, Action <int>, Animation <int>, Comedy <int>, Drama <int>,
## #   Documentary <int>, Romance <int>, Short <int>
```

```r
str(movies)
```

```
## tibble [58,788 x 24] (S3: tbl_df/tbl/data.frame)
##  $ title       : chr [1:58788] "$" "$1000 a Touchdown" "$21 a Day Once a Month" "$40,000" ...
##  $ year        : int [1:58788] 1971 1939 1941 1996 1975 2000 2002 2002 1987 1917 ...
##  $ length      : int [1:58788] 121 71 7 70 71 91 93 25 97 61 ...
##  $ budget      : int [1:58788] NA NA NA NA NA NA NA NA NA NA ...
##  $ rating      : num [1:58788] 6.4 6 8.2 8.2 3.4 4.3 5.3 6.7 6.6 6 ...
##  $ votes       : int [1:58788] 348 20 5 6 17 45 200 24 18 51 ...
##  $ r1          : num [1:58788] 4.5 0 0 14.5 24.5 4.5 4.5 4.5 4.5 4.5 ...
##  $ r2          : num [1:58788] 4.5 14.5 0 0 4.5 4.5 0 4.5 4.5 0 ...
##  $ r3          : num [1:58788] 4.5 4.5 0 0 0 4.5 4.5 4.5 4.5 4.5 ...
##  $ r4          : num [1:58788] 4.5 24.5 0 0 14.5 14.5 4.5 4.5 0 4.5 ...
##  $ r5          : num [1:58788] 14.5 14.5 0 0 14.5 14.5 24.5 4.5 0 4.5 ...
##  $ r6          : num [1:58788] 24.5 14.5 24.5 0 4.5 14.5 24.5 14.5 0 44.5 ...
##  $ r7          : num [1:58788] 24.5 14.5 0 0 0 4.5 14.5 14.5 34.5 14.5 ...
##  $ r8          : num [1:58788] 14.5 4.5 44.5 0 0 4.5 4.5 14.5 14.5 4.5 ...
##  $ r9          : num [1:58788] 4.5 4.5 24.5 34.5 0 14.5 4.5 4.5 4.5 4.5 ...
##  $ r10         : num [1:58788] 4.5 14.5 24.5 45.5 24.5 14.5 14.5 14.5 24.5 4.5 ...
##  $ mpaa        : chr [1:58788] "" "" "" "" ...
##  $ Action      : int [1:58788] 0 0 0 0 0 0 1 0 0 0 ...
##  $ Animation   : int [1:58788] 0 0 1 0 0 0 0 0 0 0 ...
##  $ Comedy      : int [1:58788] 1 1 0 1 0 0 0 0 0 0 ...
##  $ Drama       : int [1:58788] 1 0 0 0 0 1 1 0 1 0 ...
##  $ Documentary: int [1:58788] 0 0 0 0 0 0 0 1 0 0 ...
##  $ Romance     : int [1:58788] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Short       : int [1:58788] 0 0 1 0 0 0 0 1 0 0 ...
```

What is the range of years of production of the movies of this data set?

```r
range(movies$year)
```

```
## [1] 1893 2005
```

What proportion of movies have their budget included in this data base, and what proportion doesn't? What are the top 5 most expensive movies in this data set?

```r
movies.with.budget = sum(!is.na(movies$budget))
movies.without.budget = sum(is.na(movies$budget))
total.movies = nrow(movies)
budget.proportions = movies.with.budget / total.movies
nobudget.proportions = movies.without.budget / total.movies

top.5.expensive = head(movies[order(-movies$budget), ], 5)
```

```r
cat("Proportion of movies that have their budget included:", round(budget.proportions, 3))
```

```
## Proportion of movies that have their budget included: 0.089
```

```r
cat("\n\nProportion of movies that do not have their budget included:", round(nobudget.proportions, 3))
```

```
##
##
## Proportion of movies that do not have their budget included: 0.911
```

```r
cat("\n\nTop 5 Most Expensive Movies:\n")
```

```
##
##
## Top 5 Most Expensive Movies:
```

```r
for (title in top.5.expensive$title) {
  cat("• ", title, "\n", sep = "")
}
```

```
## • Spider-Man 2
## • Titanic
## • Troy
## • Terminator 3: Rise of the Machines
## • Waterworld
```

What are the top 5 longest movies?

```r
top.5.longest = head(movies[order(-movies$length), ], 5)
cat("\nTop 5 Most Longest Movies:\n")
```

```
##
## Top 5 Most Longest Movies:
```

```r
for (title in top.5.longest$title) {
  cat("• ", title, "\n", sep = "")
}
```

```
## • Cure for Insomnia, The
## • Longest Most Meaningless Movie in the World, The
## • Four Stars
## • Resan
## • Out 1
```

Of all short movies, which one is the shortest (in minutes)? Which one is the longest? How long are the shortest and the longest short movies?

```
short.movies = movies[movies$Short == 1, ]

shortest.short = which.min(short.movies$length)
shortest.short.title = as.character(short.movies$title[shortest.short])
shortest.short.time = min(short.movies$length)

longest.short = which.max(short.movies$length)
longest.short.title = as.character(short.movies$title[longest.short])
longest.short.time = max(short.movies$length)

cat('Shortest Movie:', shortest.short.title)
```

```
## Shortest Movie: 17 Seconds to Sophie
```

```
cat('\nRun Time:', shortest.short.time, 'minute')
```

```
##
## Run Time: 1 minute
```

```
cat('\n\nLongest Movie:', longest.short.title)
```

```
##
##
## Longest Movie: 10 jaar leuven kort
```

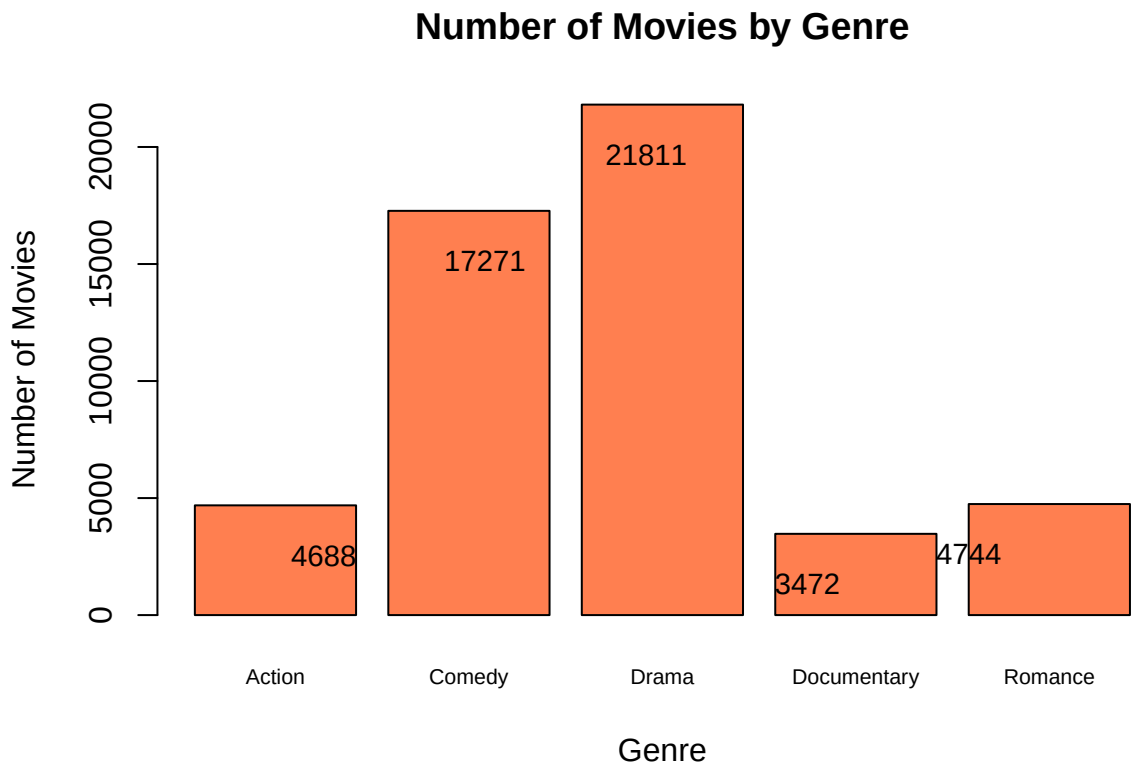
```
cat('\nRun Time', longest.short.time, 'minutes')
```

```
##
## Run Time 240 minutes
```

How many movies of each genre (action, animation, comedy, drama, documentary, romance, short) are there in this data base?

```
Filtered.movies = subset(movies, Action == 1 | Comedy == 1 |
                           Drama == 1 | Documentary == 1 | Romance == 1)
Genre.count = c(
  Action = sum(Filtered.movies$Action),
  Comedy = sum(Filtered.movies$Comedy),
  Drama = sum(Filtered.movies$Drama),
  Documentary = sum(Filtered.movies$Documentary),
  Romance = sum(Filtered.movies$Romance)
```

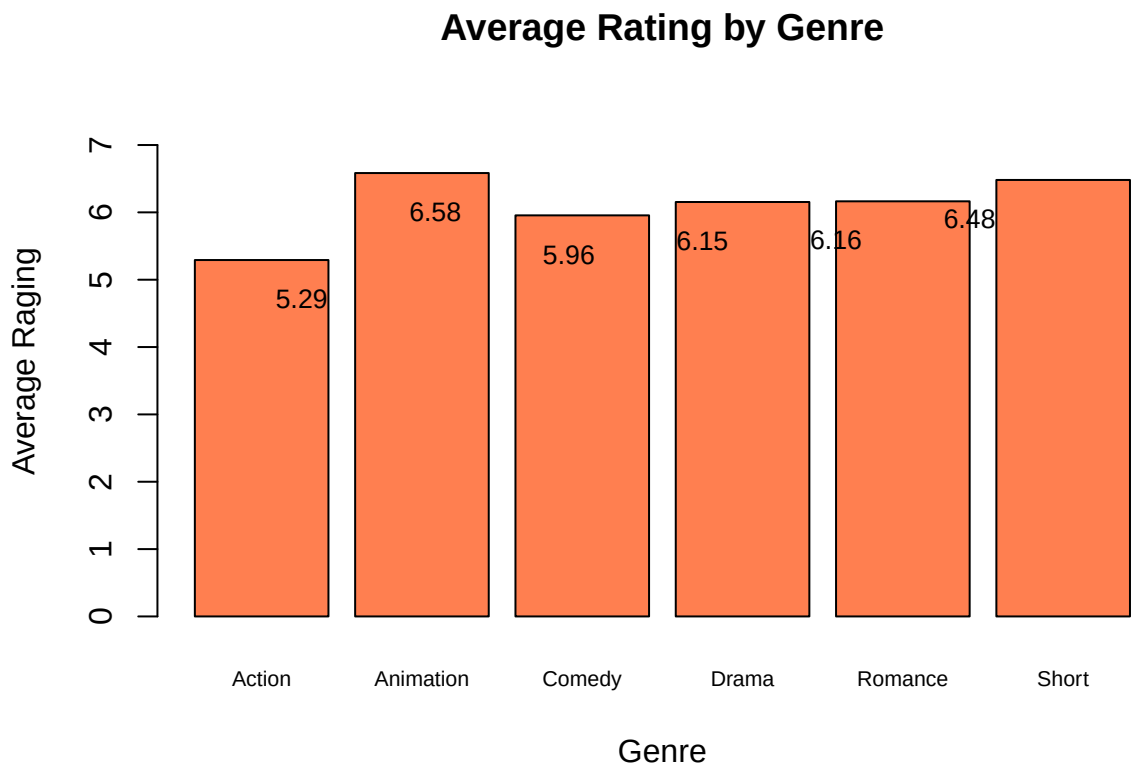
```
)

barplot(Genre.count, main = 'Number of Movies by Genre',
        xlab = 'Genre', ylab = 'Number of Movies',
        col = 'coral',
        ylim = c(0, max(Genre.count) + 10), cex.names = 0.67)
text(x = 1:length(Genre.count) , y = Genre.count - 1000,
     labels = Genre.count, cex = 0.9, col = "black", pos = 1)
```



What is the average rating of all movies within each genre?

```
Action.avg.rating = mean((movies %>% filter(Action == 1))$rating, na.rm = TRUE)
Animation.avg.rating = mean((movies %>% filter(Animation == 1))$rating, na.rm = TRUE)
Comedy.avg.rating = mean((movies %>% filter(Comedy == 1))$rating, na.rm = TRUE)
Drama.avg.rating = mean((movies %>% filter(Drama == 1))$rating, na.rm = TRUE)
Romance.avg.rating = mean((movies %>% filter(Romance == 1))$rating, na.rm = TRUE)
Short.avg.rating = mean((movies %>% filter(Short == 1))$rating, na.rm = TRUE)

avg.genre.ratings = c(Action = Action.avg.rating,
                      Animation = Animation.avg.rating,
                      Comedy = Comedy.avg.rating,
                      Drama = Drama.avg.rating,
```

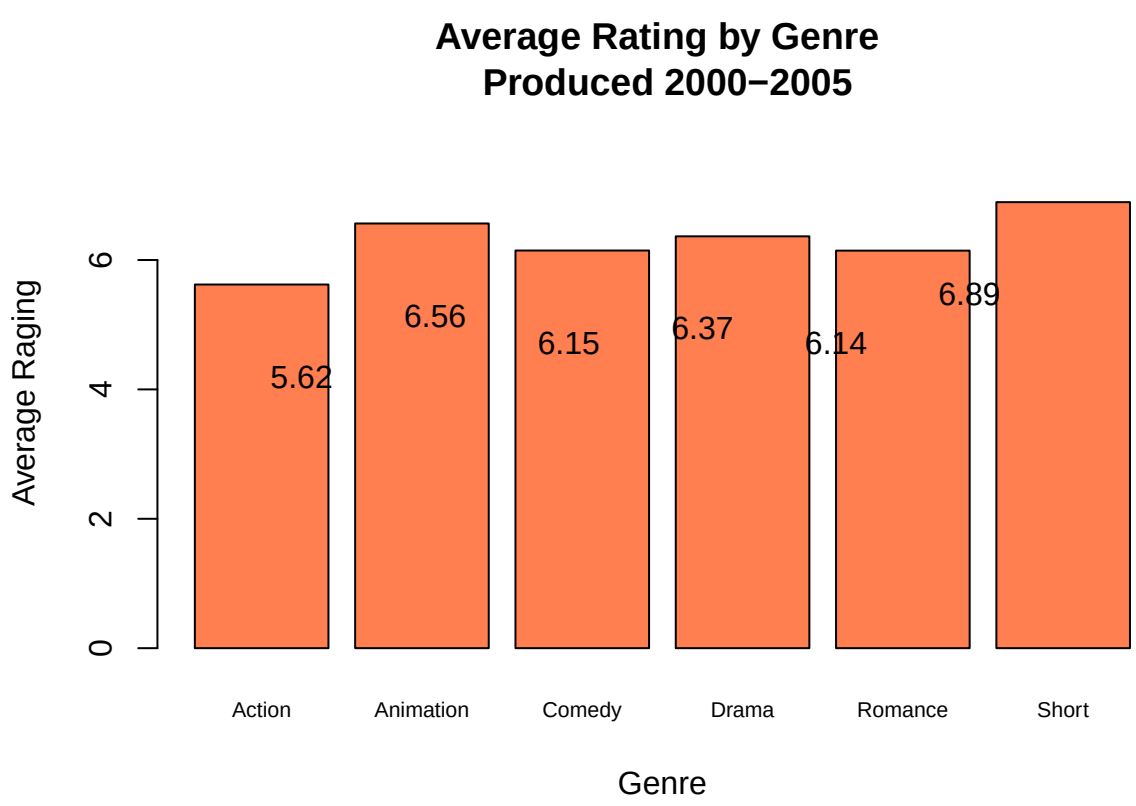
```
                      Romance = Romance.avg.rating,
                      Short = Short.avg.rating)

barplot(avg.genre.ratings,
        main = 'Average Rating by Genre', xlab = 'Genre', ylab = 'Average Raging',
        col = 'coral', ylim = c(0, max(avg.genre.ratings) + 1),
        names.arg = names(avg.genre.ratings), cex.names = 0.67)
text(x = 1:length(avg.genre.ratings) , y = avg.genre.ratings - 1,
     labels = round(avg.genre.ratings, 2), cex = 0.8, col = "black", pos = 3)
```



What is the average rating of all movies within each genre that were produced in the years 2000-2005?

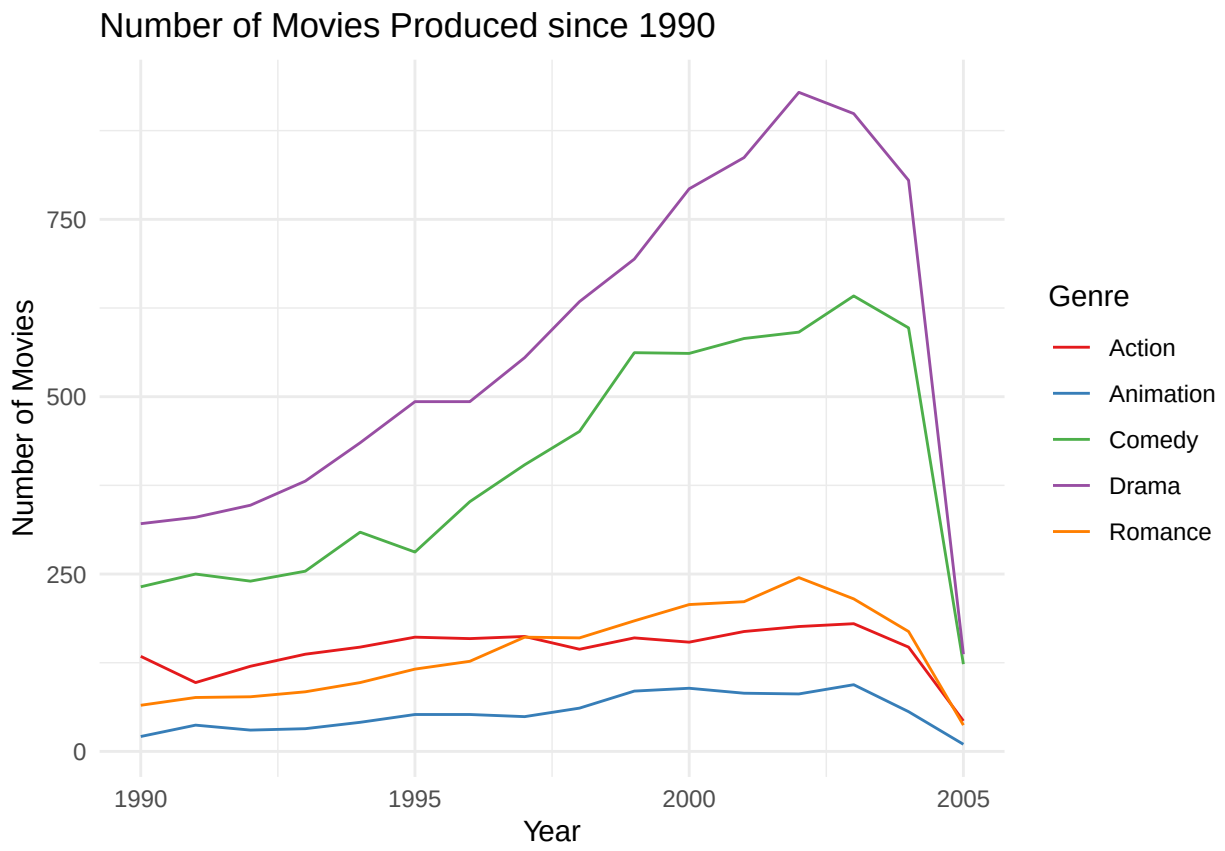
```
Action.avg.rating.7 = mean((movies %>% filter(Action == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)

Animation.avg.rating.7 = mean((movies %>% filter(Animation == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)

Comedy.avg.rating.7 = mean((movies %>% filter(Comedy == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)
```

```r
Drama.avg.rating.7 = mean((movies %>% filter(Drama == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)

Romance.avg.rating.7 = mean((movies %>% filter(Romance == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)

Short.avg.rating.7 = mean((movies %>% filter(Short == 1 &
2000 <= year & year <= 2005))$rating, na.rm = TRUE)


avg.genre.ratings.7 = c(Action = Action.avg.rating.7,
                        Animation = Animation.avg.rating.7,
                        Comedy = Comedy.avg.rating.7,
                        Drama = Drama.avg.rating.7,
                        Romance = Romance.avg.rating.7,
                        Short = Short.avg.rating.7)

barplot(avg.genre.ratings.7,
        main = 'Average Rating by Genre \n Produced 2000-2005',
        xlab = 'Genre', ylab = 'Average Raging',
        col = 'coral', ylim = c(0, max(avg.genre.ratings.7) + 1),
        names.arg = names(avg.genre.ratings.7), cex.names = 0.67)
text(x = 1:length(avg.genre.ratings.7) , y = avg.genre.ratings.7 - 1,
     labels = round(avg.genre.ratings.7, 2), cex = 0.99, col = "black", pos = 1)
```



Movies Produced, 1990 - 2005

```
Action.avg.rating.8 = movies %>% filter(Action == 1 &
1990 <= year & year <= max(year, na.rm = TRUE))

Animation.avg.rating.8 = movies %>% filter(Animation == 1 &
1990 <= year & year <= max(year, na.rm = TRUE))

Comedy.avg.rating.8 = movies %>% filter(Comedy == 1 &
1990 <= year & year <= max(year, na.rm = TRUE))

Drama.avg.rating.8 = movies %>% filter(Drama == 1 &
1990 <= year & year <= max(year, na.rm = TRUE))

Romance.avg.rating.8 = movies %>% filter(Romance == 1 &
1990 <= year & year <= max(year, na.rm = TRUE))

Genre.movies.time = bind_rows(
                      data.frame(Genre = 'Action', year = Action.avg.rating.8$year),
                      data.frame(Genre = 'Animation',year =  Animation.avg.rating.8$year),
                      data.frame(Genre = 'Comedy', year = Comedy.avg.rating.8$year),
                      data.frame(Genre = 'Drama', year = Drama.avg.rating.8$year),
                      data.frame(Genre = 'Romance', year = Romance.avg.rating.8$year)
                )

plot = ggplot(Genre.movies.time, aes(x = year, color = Genre)) +
  geom_line(stat = "count") +
  labs(title = "Number of Movies Produced since 1990",
       x = "Year", y = "Number of Movies", color = "Genre") +
  theme_minimal() +
  scale_color_brewer(palette = "Set1") +
  theme(legend.position = "right")
print(plot)
```

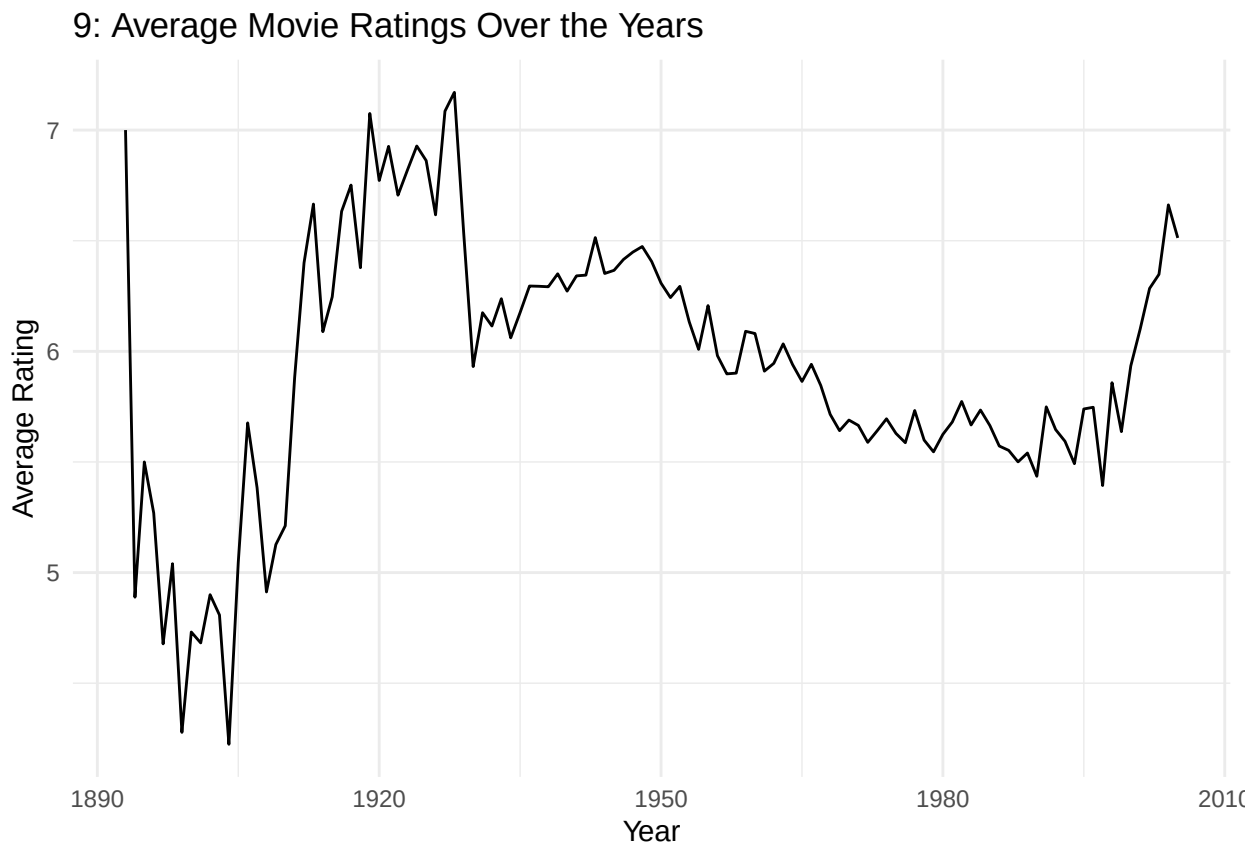


How has the popularity of movie genres changed over the decades?

By looking at the graph above, it becomes evident the gap between Comedy, Drama, and every other genre. At ~1996, Romance movies became more popular than Action films but both have declined ~2002, following the rest of the graphs pattern. Considering Comedy movies, they capped out at around ~600 movies being produced each year while Drama movies continued to grow. What is interesting is that although Drama movies continued to grow, them alongside all genres began to decline at ~2002 indicating an obstacle in the film industry. Further research is required to understand how this obstacle played a role.

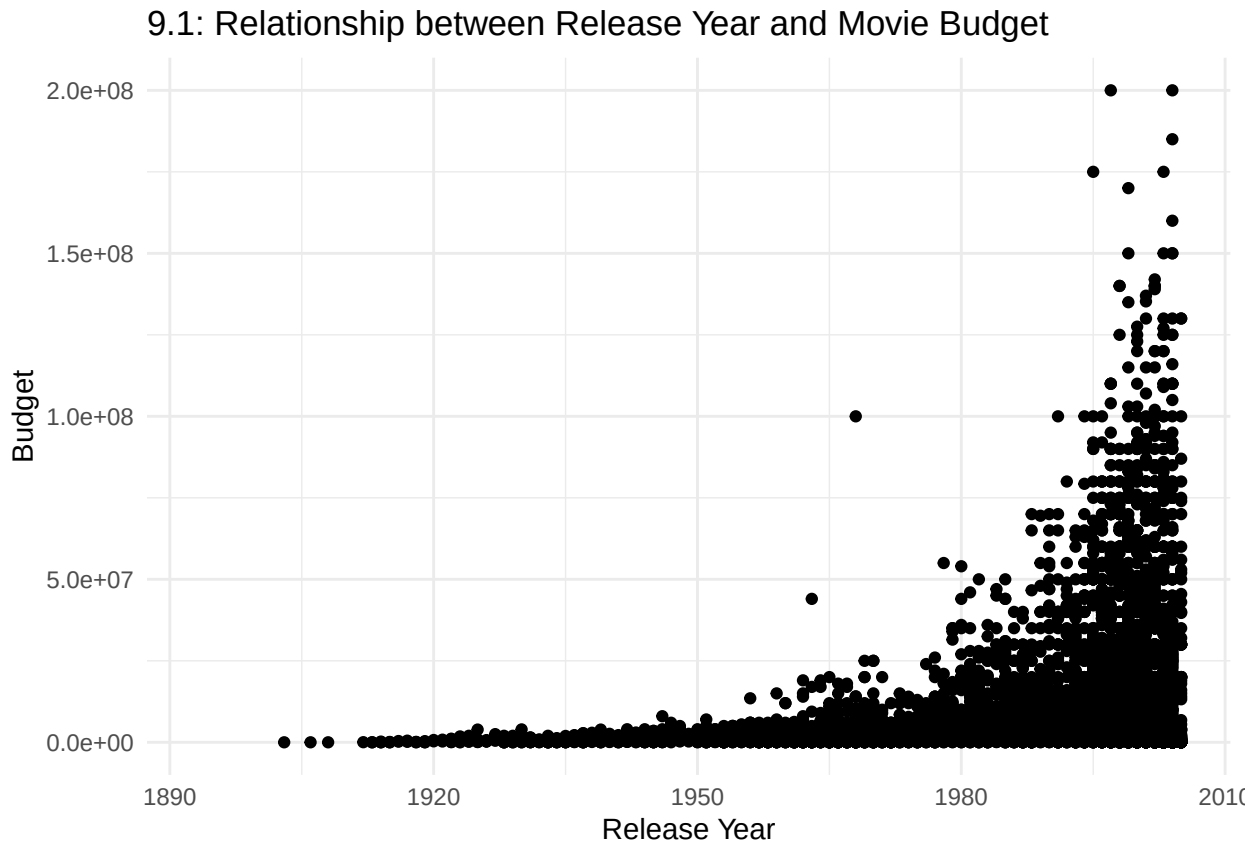
What is the trend of average movie ratings over the years?

```
avg.ratings.by.year = movies %>%
  group_by(year) %>%
  summarize(avg.rating = mean(rating, na.rm = TRUE))

plot.9 = ggplot(avg.ratings.by.year, aes(x = year, y = avg.rating)) +
  geom_line() +
  labs(title = "9: Average Movie Ratings Over the Years",
       x = "Year",
       y = "Average Rating") +
  theme_minimal()
print(plot.9)
```




```
plot.9.1 = ggplot(movies, aes(x = year, y = budget)) +
  geom_point() +
  labs(title = "9.1: Relationship between Release Year and Movie Budget",
       x = "Release Year",
       y = "Budget") +
  theme_minimal()
print(plot.9.1)
```

```
## Warning: Removed 53573 rows containing missing values (‘geom_point()‘).
```



As seen through graph 9, viewer sentiment was mixed and the data could be viewed as unreliable due to its extremities. Up until ~1930, the average movie rating was decreasing for an unknown reason that will consider further research. My guess is that this is when the film industry was just beginning to develop and new ideas were emerging/being tested. However, ~1990 average ratings began to plateau marking a critical point since average ratings increased from here. When comparing this with the budget for movies' release years, it is noticed that budgets increased exponentially around this same time. Therefore, it can be concluded that these increased budgets have caused average movie ratings to rise.

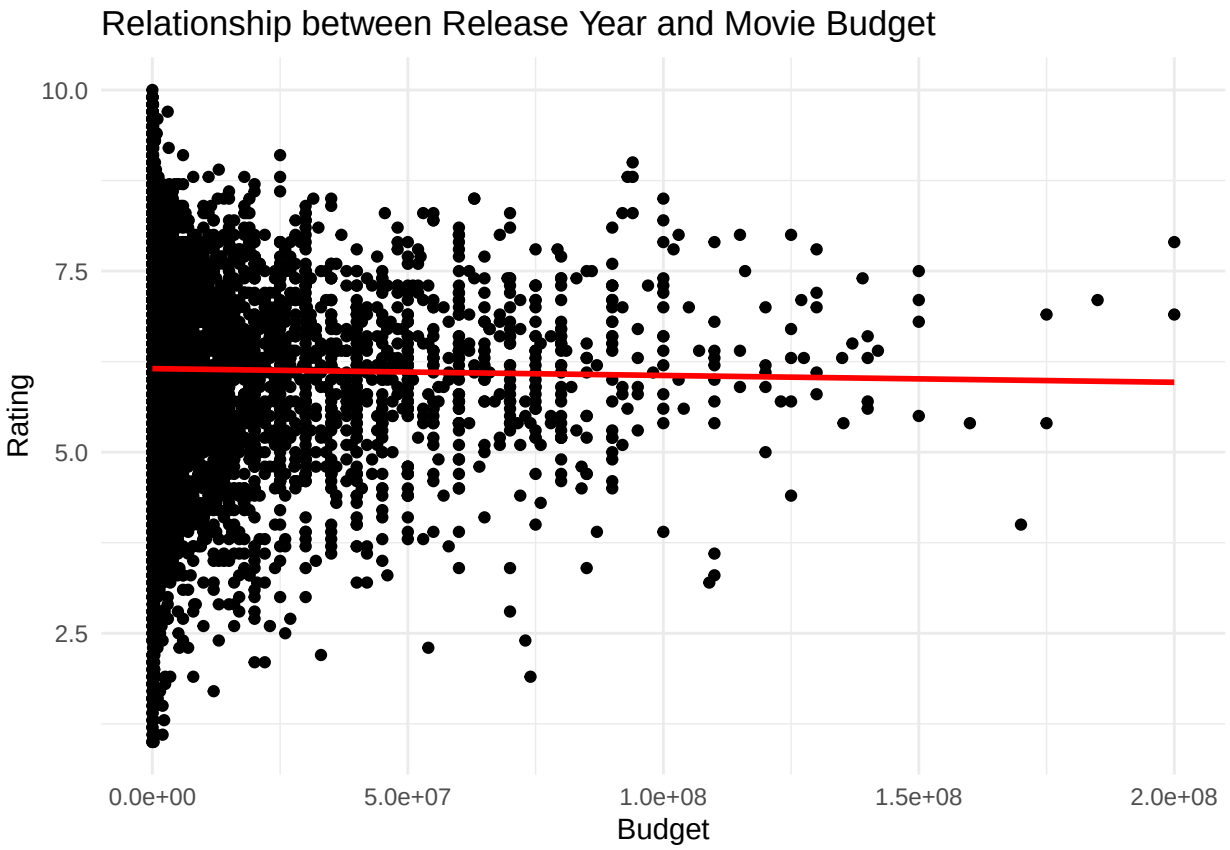
Is there a relationship between budgets and ratings?

```
plot.10 = ggplot(movies, aes(x = budget, y = rating)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE, color = 'red') +
  labs(title = "Relationship between Release Year and Movie Budget",
       x = "Budget",
       y = "Rating") +
  theme_minimal()
print(plot.10)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 53573 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 53573 rows containing missing values (`geom_point()`).
```



Yes! By comparing the budgets of movies to their rating, we notice through our trend line that as budgets increase, movies tend to have lower ratrings. However, this decrease in rating is not significant since our slope is very close to being neutral. If we look at the scatter plot itself, we notice that ratings cluster much more around the 7 are meaning that a big budget is not requierd to achieve this rating. Furthermore, it is noticed that the highest ratings are on the lower half of the budget, indicating that a movie does not need a budget of $100,000,000 to achieve a high rating.