Exporing Data in the The Internet Movie Database

Jasmine Riggs

3/15/2020

Introduction

In this project, I explore the `movies` data set from the `ggplot2movies` package. The data is from the internet movie database, imdb.com. It's one of the largest movie databases today and is a collection of movie data supplied by studios and fans. There are 58,788 rows and 24 variables in this data set.

Exploratory Data Analysis

I develop some questions that come to mind and that I think are interesting to know. I also keep in mind what kind of visualizations I can create based upon these questions and answers.

What is the range of years of production of the movies of this data set (i.e. what is the year of production of the oldest movie and of the most recent movie in this data set)?

```
min(movies$year) #year of the oldest movie
```

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

```
max(movies$year) #year of the most recent movie
```

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

The year of production of the oldest movie is 1893 and the year of production of the most recent movie in this data set is 2005.

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

```
budgetNAs <- sum(is.na(movies$budget)) #number of NAs in the budget column
totalMovies <- nrow(movies) #total number of rows (movies)
budgetNAs/totalMovies #proportion of movies w/ NO budget included
```

```
## [1] 0.9112914
```

```
1-budgetNAs/totalMovies #proportion of movies w/ budget included
```

```
## [1] 0.08870858
```

```
#top 5 most expensive movies
movies[head(order(movies$budget,decreasing = T),5),c("title","budget")]
```

```
##                                      title    budget
## 48518                        Spider-Man 2 200000000
## 52348                             Titanic 200000000
## 53437                                Troy 185000000
## 51244 Terminator 3: Rise of the Machines 175000000
## 56212                           Waterworld 175000000
```

About 91.1% of movies have their budget included in this data set and the rest, 8.9%, do not. The top 5 most expensive movies in this data set are "Spider-Man 2", "Titanic", "Troy", "Terminator 3: Rise of the Machines", and "Waterworld".

What are top 5 longest movies?

```
#top 5 longest movies
movies[head(order(movies$length,decreasing = T),5),c("title","length")]
```

```
##                                                 title length
## 11937                          Cure for Insomnia, The   5220
## 30574 Longest Most Meaningless Movie in the World, The   2880
## 18741                                     Four Stars   1100
## 42957                                          Resan    873
## 38435                                          Out 1    773
```

The top 5 longest movies are The Cure for Insomnia, The Longest Most Meaningless Movie in the World, Four Stars, Resan, and 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?

```
moviesShort <- movies[movies$Short==1,] #filtered data with short films only

#shortest short film
moviesShort[head(order(moviesShort$length),1),c("title","length")]
```

```
##                     title length
## 206 17 Seconds to Sophie      1
```

```
#number of short films w/ length of 1 minute
sum(moviesShort$length==1)
```

```
## [1] 165
```

```
#longest short film
moviesShort[head(order(moviesShort$length,decreasing = T),1),c("title","length")]
```

```
##                    title length
## 115 10 jaar leuven kort    240
```

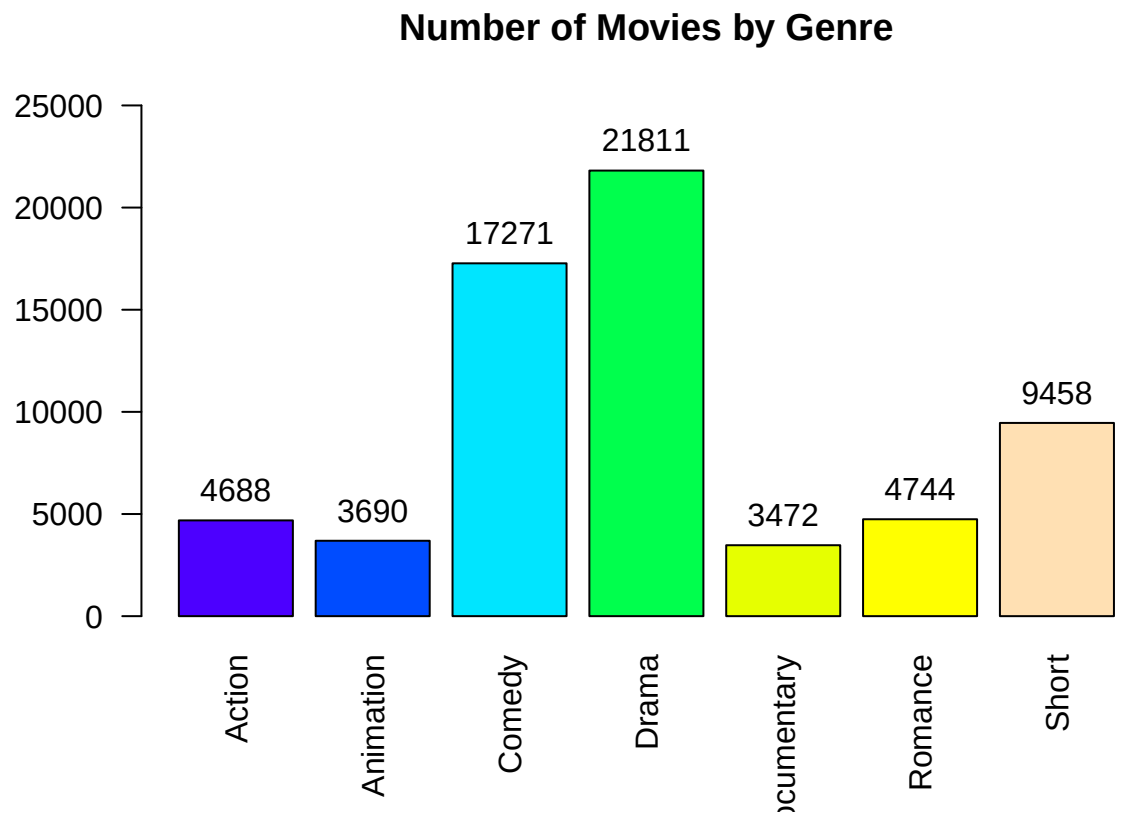
The shortest short film is 17 Seconds to Sophie with a length of 1 minute. However, I noticed that there are a total of 165 short films with a length of 1 minute. 17 Seconds to Sophie just happens to be at the top of the list because it's in alphabetical order. The longest short film is 10 jaar leuven kort with a length of 240 minutes (which could arguably not be a short film).

How many movies of each genre are there in this data set?

```
#create new aggregated data frame by summing binary genre columns
moviesGenre <- data.frame("Genre" = c("Action","Animation","Comedy","Drama",
                                      "Documentary","Romance","Short"),
                          "Count" = c(sum(movies$Action),sum(movies$Animation),
                                      sum(movies$Comedy),sum(movies$Drama),
                                      sum(movies$Documentary),sum(movies$Romance),
                                      sum(movies$Short)))
moviesGenre
```

```
##         Genre Count
## 1      Action  4688
## 2   Animation  3690
## 3      Comedy 17271
## 4       Drama 21811
## 5 Documentary  3472
## 6     Romance  4744
## 7       Short  9458
```

```
#create barplot
xx <- barplot(height = moviesGenre$Count,
              names.arg = moviesGenre$Genre,
              las = 2,
              col = topo.colors(length(moviesGenre$Genre)),
              ylim = c(0,25000),
              main = "Number of Movies by Genre")
#add data labels on top of each bar
text(x = xx, y = moviesGenre$Count, label = moviesGenre$Count, pos = 3)
```

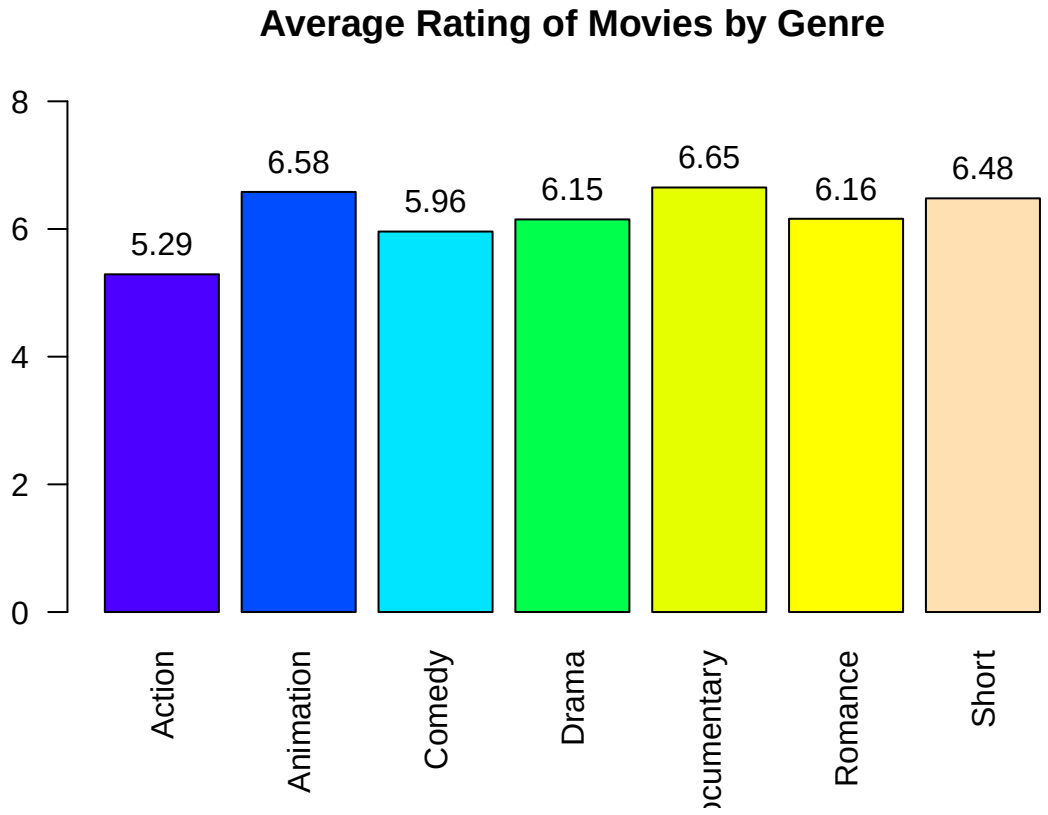


What is the average rating of all movies within each genre?

```
#create new aggregated data frame by averaging rating column if value in each
#genre column is equal to 1
moviesRating <- data.frame("Genre" = c("Action","Animation","Comedy","Drama",
                                       "Documentary","Romance","Short"),
                           "Avg Rating" = c(mean(movies[movies$Action==1,]$rating),
                                            mean(movies[movies$Animation==1,]$rating),
                                            mean(movies[movies$Comedy==1,]$rating),
                                            mean(movies[movies$Drama==1,]$rating),
                                            mean(movies[movies$Documentary==1,]$rating),
                                            mean(movies[movies$Romance==1,]$rating),
                                            mean(movies[movies$Short==1,]$rating)))
#round 2 decimal places
moviesRating$Avg.Rating <- round(moviesRating$Avg.Rating,2)
moviesRating
```

```
##         Genre Avg.Rating
## 1      Action       5.29
## 2   Animation       6.58
## 3      Comedy       5.96
## 4       Drama       6.15
## 5 Documentary       6.65
## 6     Romance       6.16
## 7       Short       6.48
```

```r
#create barplot
yy <- barplot(height = moviesRating$Avg.Rating,
              names.arg = moviesRating$Genre,
              las = 2,
              col = topo.colors(length(moviesRating$Genre)),
              ylim = c(0,8),
              main = "Average Rating of Movies by Genre")
#add data labels on top of each bar
text(x = yy, y = moviesRating$Avg.Rating, label = moviesRating$Avg.Rating, pos = 3)
```



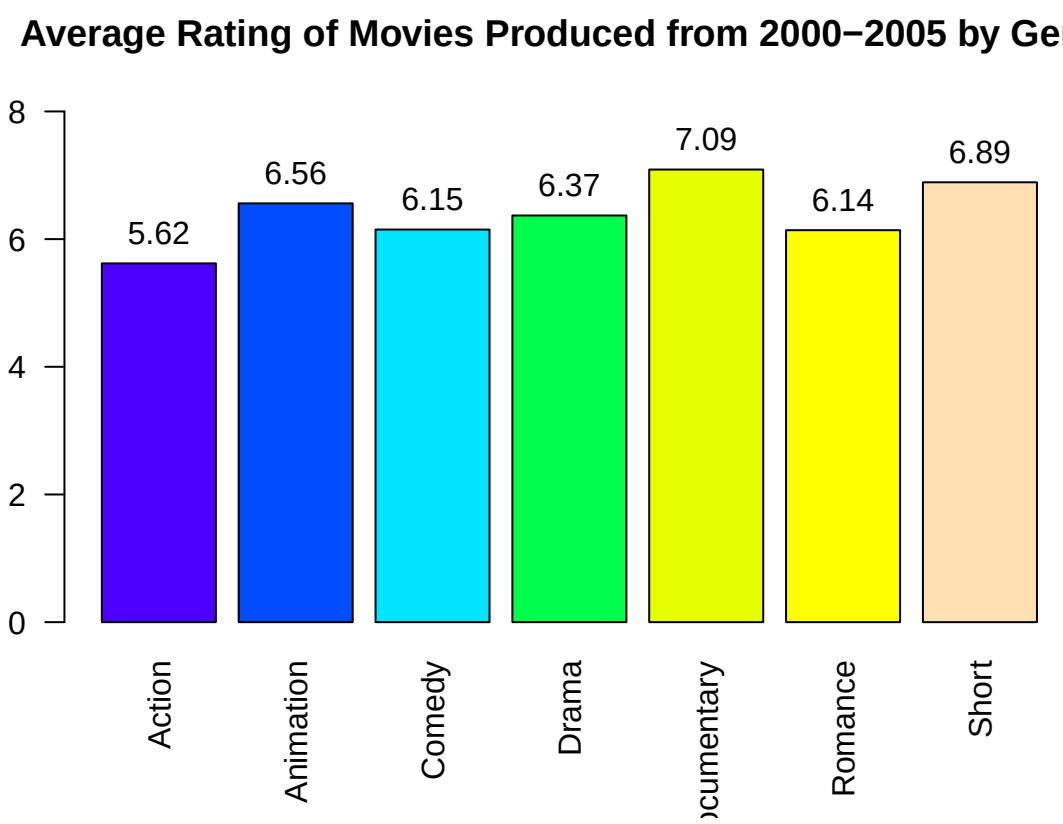
What is the average rating of all movies within each genre that were produced in the years 2000-2005?

```r
#create new aggregated data frame by averaging rating column if value in each
#genre column is equal to 1
moviesRating2000_2005 <- data.frame("Genre" = c("Action","Animation","Comedy","Drama",
                                    "Documentary","Romance","Short"),
                         "Avg Rating" = c(mean(movies[movies$Action==1
                                                      & movies$year>=2000
                                                      & movies$year<="2005",]$rating),
                                          mean(movies[movies$Animation==1
                                                      & movies$year>=2000
                                                      & movies$year<="2005",]$rating),
                                          mean(movies[movies$Comedy==1
                                                      & movies$year>=2000
                                                      & movies$year<="2005",]$rating),
                                          mean(movies[movies$Drama==1
```

```
                                                    & movies$year>=2000
                                                    & movies$year<="2005",]$rating),
                                            mean(movies[movies$Documentary==1
                                                    & movies$year>=2000
                                                    & movies$year<="2005",]$rating),
                                            mean(movies[movies$Romance==1
                                                    & movies$year>=2000
                                                    & movies$year<="2005",]$rating),
                                            mean(movies[movies$Short==1
                                                    & movies$year>=2000
                                                    & movies$year<="2005",]$rating)))
#round 2 decimal places
moviesRating2000_2005$Avg.Rating <- round(moviesRating2000_2005$Avg.Rating,2)
moviesRating2000_2005
```

```
##         Genre Avg.Rating
## 1      Action       5.62
## 2   Animation       6.56
## 3      Comedy       6.15
## 4       Drama       6.37
## 5 Documentary       7.09
## 6     Romance       6.14
## 7       Short       6.89
```

```
#create barplot
zz <- barplot(height = moviesRating2000_2005$Avg.Rating,
              names.arg = moviesRating2000_2005$Genre,
              las = 2,
              col = topo.colors(length(moviesRating2000_2005$Genre)),
              ylim = c(0,8),
              main = "Average Rating of Movies Produced from 2000-2005 by Genre")
#add data labels on top of each bar
text(x = zz, y = moviesRating2000_2005$Avg.Rating,
     label = moviesRating2000_2005$Avg.Rating, pos = 3)
```



For this part, I wanted to combine multiple plots on one graph. For each genre (not including short films) plot movies from 1990-2005 and plot a function of the number of movies in this data set of corresponding genre produced by year, for years from 1990 until the last year recorded.

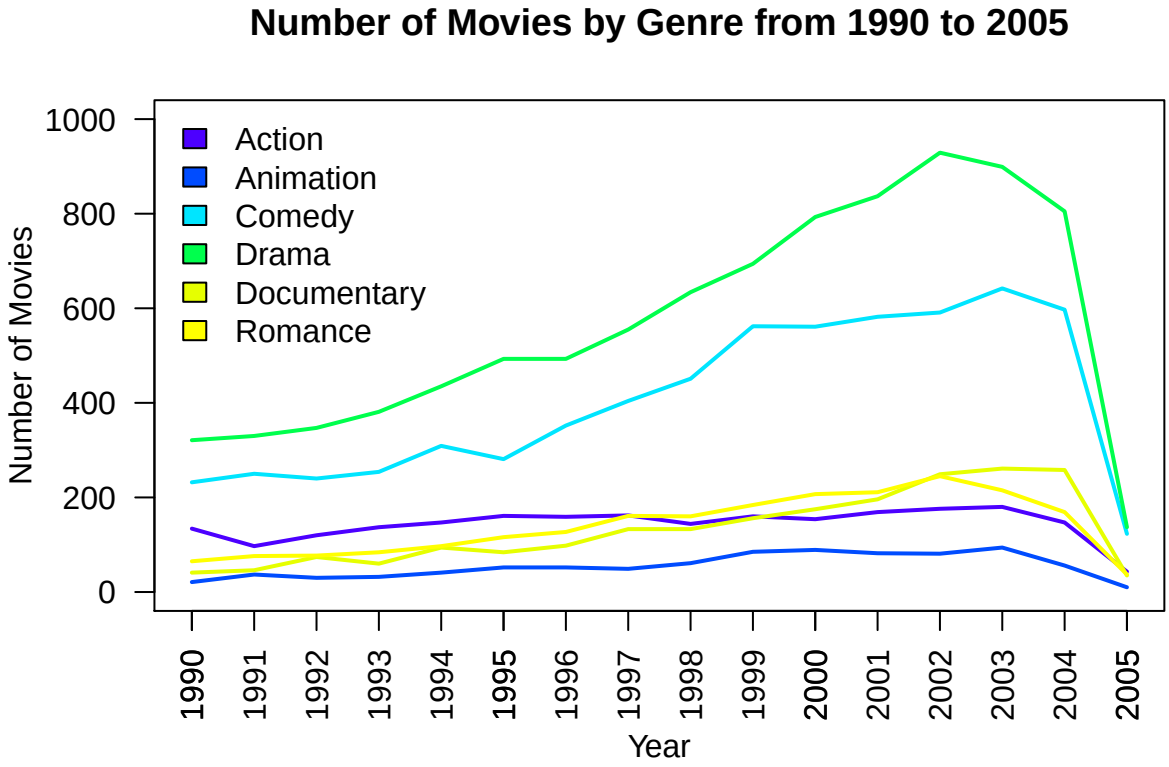
```
#create empty dataframe with first column as years
movies1990_2005 <- data.frame("Year" = 1990:2005)
#add genres as column names
movies1990_2005[c("Action","Animation","Comedy","Drama",
                  "Documentary","Romance")] <- NA
#for each genre column, loop through each year and add up the number of movies
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Action)
  movies1990_2005[yr-1989,"Action"] <- count
}
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Animation)
  movies1990_2005[yr-1989,"Animation"] <- count
}
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Comedy)
  movies1990_2005[yr-1989,"Comedy"] <- count
}
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Drama)
  movies1990_2005[yr-1989,"Drama"] <- count
}
```

```
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Documentary)
  movies1990_2005[yr-1989,"Documentary"] <- count
}
for (yr in 1990:2005) {
  count <- sum(movies[movies$year==yr,]$Romance)
  movies1990_2005[yr-1989,"Romance"] <- count
}
movies1990_2005
```

```
##    Year Action Animation Comedy Drama Documentary Romance
## 1  1990    134        21    232   321          41      65
## 2  1991     97        37    250   330          46      76
## 3  1992    120        30    240   347          74      77
## 4  1993    137        32    254   381          60      84
## 5  1994    147        41    309   435          94      97
## 6  1995    161        52    281   493          84     116
## 7  1996    159        52    352   493          98     127
## 8  1997    162        49    404   555         133     161
## 9  1998    144        61    451   634         133     160
## 10 1999    160        85    562   694         156     184
## 11 2000    154        89    561   793         175     207
## 12 2001    169        82    582   837         196     211
## 13 2002    176        81    591   929         249     245
## 14 2003    180        94    642   899         261     215
## 15 2004    147        56    597   805         258     169
## 16 2005     43        10    123   137          35      37
```

```
#plot
plot(1, type="n", xlab="Year", ylab="Number of Movies",
     xlim=c(1990,2005), ylim=c(0, 1000), las = 2,
     main = "Number of Movies by Genre from 1990 to 2005")
axis(1, at=seq(1990, 2005, 1), las = 2)
lines(movies1990_2005$Year,movies1990_2005$Action,
      col = topo.colors(length(moviesRating2000_2005$Genre))[1],
      lwd = 2)
lines(movies1990_2005$Year,movies1990_2005$Animation,
      col = topo.colors(length(moviesRating2000_2005$Genre))[2],
      lwd = 2)
lines(movies1990_2005$Year,movies1990_2005$Comedy,
      col = topo.colors(length(moviesRating2000_2005$Genre))[3],
      lwd = 2)
lines(movies1990_2005$Year,movies1990_2005$Drama,
      col = topo.colors(length(moviesRating2000_2005$Genre))[4],
      lwd = 2)
lines(movies1990_2005$Year,movies1990_2005$Documentary,
      col = topo.colors(length(moviesRating2000_2005$Genre))[5],
      lwd = 2)
lines(movies1990_2005$Year,movies1990_2005$Romance,
      col = topo.colors(length(moviesRating2000_2005$Genre))[6],
      lwd = 2)
#legend
legend("topleft",legend=colnames(movies1990_2005)[2:7],
```

```
       fill=topo.colors(length(moviesRating2000_2005$Genre)),bty='n')
```

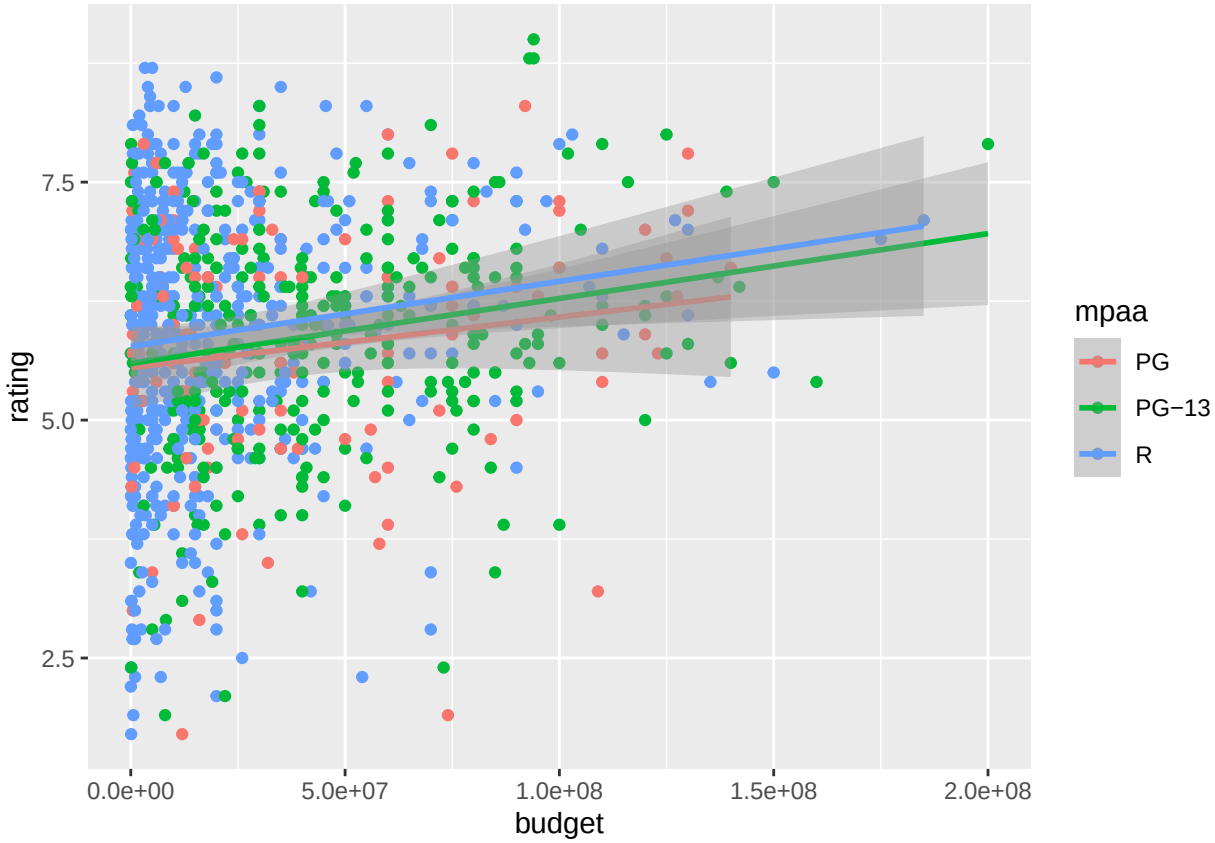


For this last section, I utilize ggplot to visulize the correlation between variables.

For movies that are not short films produced from 2000-2005, analyze the correlation between budget and rating by plotting budget on the y-axis and rating on the x-axis.

```
#create dataframe omitting rows where budget is NA
moviesBudget2000_2005 <- movies[!is.na(movies$budget) &
                                  movies$budget!=0 &
                                  movies$year>=2000 &
                                  movies$year<=2005 &
                                  movies$Short==0 &
                                  movies$mpaa!="",]
#plot
library(ggplot2)
g <-ggplot(data = moviesBudget2000_2005,aes(x = budget, y = rating,
                                            color = mpaa)) + geom_point()
g + geom_smooth(method="lm", level = 0.99)
```

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



For all mpaa ratings, as budget increases, rating slightly increases as well. You can see this because all three of the regression lines are slightly positive. There are more rated R movies with lower budgets because there are way more blue points on the left side of the graph than any other color. The most expensive movie in this data set was a PG-13 movie with a rating that is quite high at about 7.9. This is shown as the green point at the far top right of the plot.