Module 3 Lab2

Juliana Perez Romero

2024-03-29

```
library(ggplot2movies)
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
```

```
library(ggplot2)
data(movies)
```

Question 1:
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)
```

```
| [1] 1893
```

```
max (movies$year)
```

```
| [1] 2005
```

Response 1:
The production of these movies spans 112 years. Beginning in 1893 and ending in 2005.

Question 2:
What proportion of movies have their budget included in this data base, and what proportion doesn't? What are top 5 most expensive movies in this data set?

```
top_5_expensive <- movies %>% arrange(desc(budget)) %>% head(5) %>% select(title,budget)
print(top_5_expensive)
```

```
| # A tibble: 5 x 2
|   title                                budget
|   <chr>                                 <int>
| 1 Spider-Man 2                       200000000
```

```
| 2 Titanic                              200000000
| 3 Troy                                 185000000
| 4 Terminator 3: Rise of the Machines 175000000
| 5 Waterworld                           175000000
```

```
movie_budget <- movies %>% summarize(budget_excluded = sum(is.na(budget)),
                                     budget_included = sum(!is.na(budget)))
print(movie_budget)
```

```
| # A tibble: 1 x 2
|   budget_excluded budget_included
|               <int>           <int>
| 1           53573            5215
```

Response 2:
53,573 movies have their budget excluded and 5,215 have their budget included. The top 5 most expensive movies in this data set are:

1. Spider-Man 2
2. Titanic
3. Troy
4. Terminator 3: Rise of the Machines
5. Waterworld

Question 3:
What are top 5 longest movies?

```
top_5_longest <- movies %>% arrange(desc(length)) %>% head(5) %>% select(title,length)
print(top_5_longest)
```

```
| # A tibble: 5 x 2
|   title                                                 length
|   <chr>                                                  <int>
| 1 Cure for Insomnia, The                                  5220
| 2 Longest Most Meaningless Movie in the World, The        2880
| 3 Four Stars                                              1100
| 4 Resan                                                    873
| 5 Out 1                                                    773
```

Response 3
The top 5 longest movies are
1. The Cure for insomnia
2. The Longest Most Meaningless Movie in the World
3. Four Stars
4. Resan

and
5. Out 1

Question 4
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?

```
#Store the movie length data column by ascending order
movie_by_length <- movies %>% arrange(length) %>% select(title, length)

#Extract the shortest movie titles where the length == minimum
shortest_movie_titles <- movie_by_length %>% filter(length == min(length)) %>% select(title)
print(shortest_movie_titles)
```

```
| # A tibble: 169 x 1
|    title
|    <chr>
|  1 17 Seconds to Sophie
|  2 2 A.M. in the Subway
|  3 Admiral Cigarette
|  4 Admiral Dewey Leading Land Parade
|  5 Alphonse and Gaston, No. 3
|  6 Ameta
|  7 Amy Muller
|  8 Arabian Gun Twirler
|  9 Arrival of McKinley's Funeral Train at Canton, Ohio
| 10 As Seen Through a Telescope
| # i 159 more rows
```

```
#Extract the longest movie titles where the length == maximum
longest_movie_titles <- movie_by_length %>% filter(length == max(length)) %>% select(title)
print(longest_movie_titles)
```

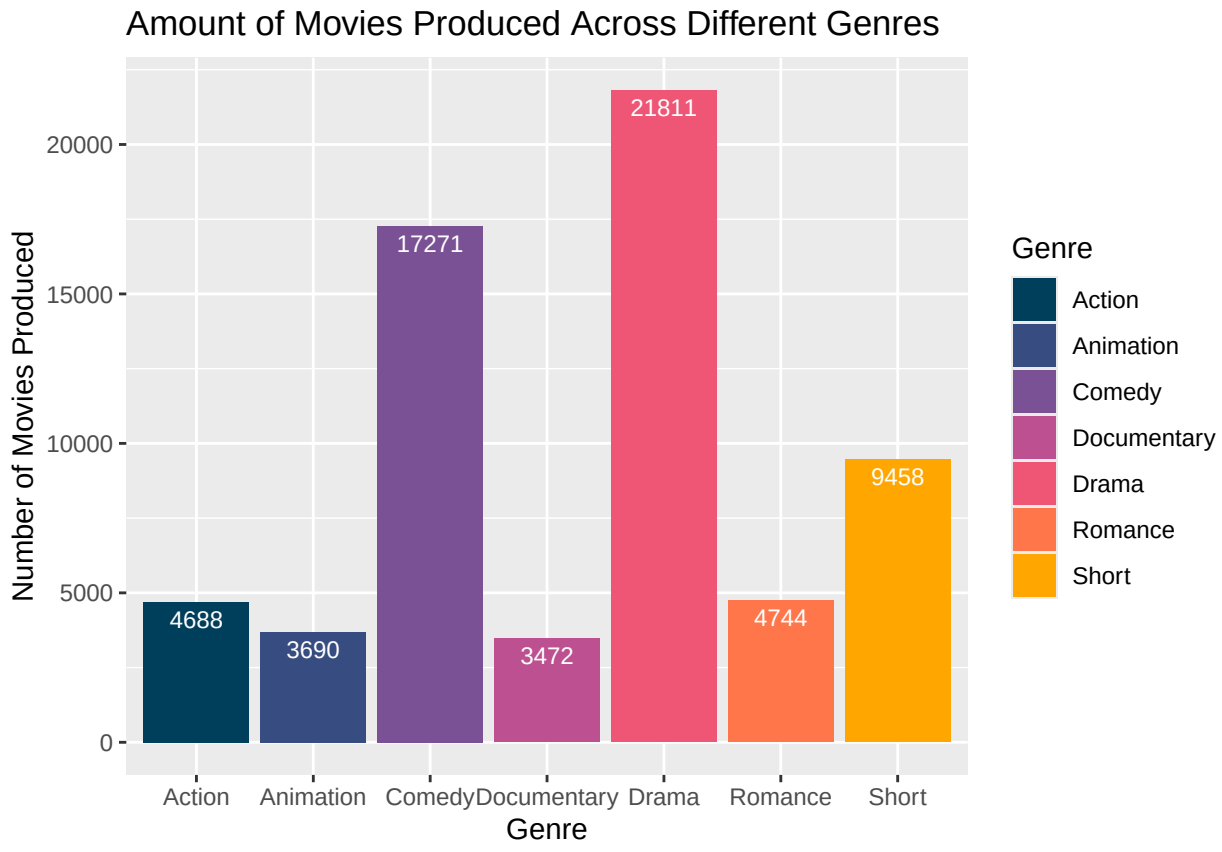
```
| # A tibble: 1 x 1
|   title
|   <chr>
| 1 Cure for Insomnia, The
```

Response 4:
The shortest length for movies in this data set is 1 minute. There are 169 movies of that length. In contrast the longest movie is 5220 minutes and the only movie of that length is The Cure for Insomnia.

Question 5:
How many movies of each genre (action, animation, comedy, drama, documentary, romance, short) are there in this data base? (use a bar plot)

```
# Calculate the sum of counts for each genre
genre_counts <- colSums(select(movies, c("Action",
                                         "Animation",
                                         "Comedy",
                                         "Drama",
                                         "Documentary",
                                         "Romance",
                                         "Short"
                                         )))

# Create a data frame from the genre counts
```

```r
movie_by_genre <- data.frame(Genre = names(genre_counts), Count = genre_counts)

# Sort the data by genre
movie_by_genre <- movie_by_genre %>%
  arrange(Genre)

#Plot bar graph
ggplot(data = movie_by_genre,
       aes( x= Genre, y = Count, fill = Genre)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Count), vjust = 1.5, size = 3, color = "white") +
  scale_fill_manual(values = c(
    "Action" = "#003f5c",
    "Animation" = "#374c80",
    "Comedy" = "#7a5195",
    "Documentary" = "#bc5090",
    "Drama" = "#ef5675",
    "Romance" = "#ff764a",
    "Short" = "#ffa600"
  )) +
  labs( title = "Amount of Movies Produced Across Different Genres",
        x = ("Genre"), y = ("Number of Movies Produced"))
```



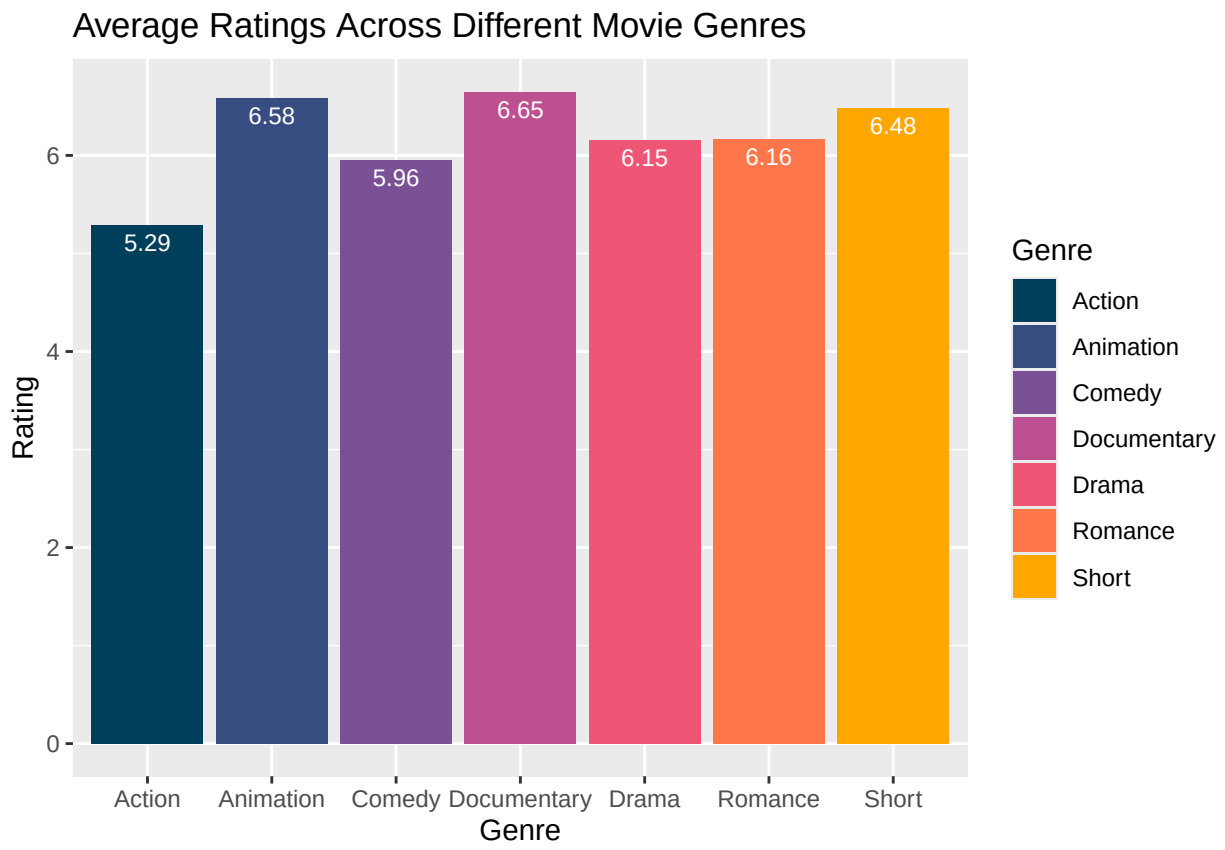
Response 5:
The following are the total number of movies in each genre:
- Action = 4,688

- Animation = 3,690
- Comedy = 17,271
- Documentary = 3,472
- Drama = 21,811
- Romance = 4,744
- Short = 9,458

Question 6:
What is the average rating of all movies within each genre? (use a bar plot)

```
action_avg <-
  movies %>%
  filter(Action ==1) %>%
  summarize(actionMeanRating = mean(rating))
animation_avg <-
  movies %>%
  filter(Animation ==1) %>%
  summarize(animationMeanRating = mean(rating))
comedy_avg <-
  movies %>%
  filter(Comedy ==1) %>%
  summarize(comedyMeanRating = mean(rating))
documentary_avg <-
  movies %>%
  filter(Documentary ==1) %>%
  summarize(documentaryMeanRating = mean(rating))
drama_avg <-
  movies %>%
  filter(Drama ==1) %>%
  summarize(dramaMeanRating = mean(rating))
romance_avg <-
  movies %>%
  filter(Romance ==1) %>%
  summarize(romanceMeanRating = mean(rating))
short_avg <-
  movies %>%
  filter(Short ==1) %>%
  summarize(shortMeanRating = mean(rating))

action_avg_df <- data.frame(Genre = "Action",
                            Rating = action_avg$actionMeanRating)
animation_avg_df <- data.frame(Genre = "Animation",
                               Rating = animation_avg$animationMeanRating)
comedy_avg_df <- data.frame(Genre = "Comedy",
                            Rating = comedy_avg$comedyMeanRating)
documentary_avg_df <- data.frame(Genre = "Documentary",
                                 Rating = documentary_avg$documentaryMeanRating)
drama_avg_df <- data.frame(Genre = "Drama",
                           Rating = drama_avg$dramaMeanRating)
romance_avg_df <- data.frame(Genre = "Romance",
                             Rating = romance_avg$romanceMeanRating)
short_avg_df <- data.frame(Genre = "Short",
                           Rating = short_avg$shortMeanRating)
```

```r
movie_genre_rating <- bind_rows(action_avg_df,
                                animation_avg_df,
                                comedy_avg_df,
                                documentary_avg_df,
                                drama_avg_df,
                                romance_avg_df,
                                short_avg_df
                                )

ggplot(data = movie_genre_rating, aes(x = Genre, y = Rating, fill = Genre)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(Rating,2)), vjust = 1.5, size = 3, color = "white") +
  scale_fill_manual(values = c(
    "Action" = "#003f5c",
    "Animation" = "#374c80",
    "Comedy" = "#7a5195",
    "Documentary" = "#bc5090",
    "Drama" = "#ef5675",
    "Romance" = "#ff764a",
    "Short" = "#ffa600"
  )) +
  labs( title = "Average Ratings Across Different Movie Genres",
        x = ("Genre"), y = ("Rating"))
```



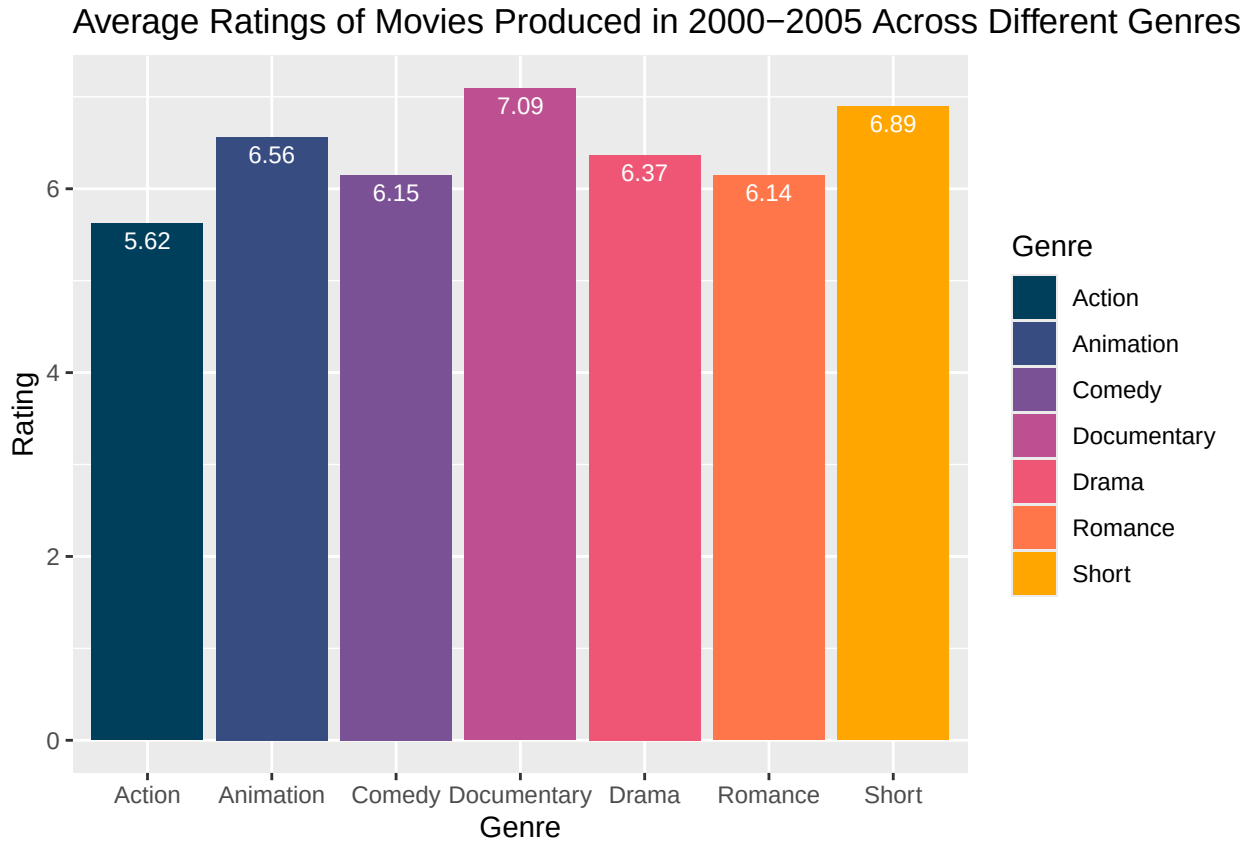
Response 6:
The average rating per genre is as approximately:

- Action: 5.29
- Animation: 6.58
- Comedy: 5.96
- Documentary: 6.65
- Drama: 6.15
- Romance: 6.16
- Short: 6.48

Question 7:
What is the average rating of all movies within each genre that were produced in the years 2000-2005? (use a bar plot)

```
action_avg_year <-
  movies %>%
  filter(Action == 1, year %in% 2000:2005) %>%
  summarize(actionMeanRating = mean(rating))
animation_avg_year <-
  movies %>%
  filter(Animation == 1, year %in% 2000:2005) %>%
  summarize(animationMeanRating = mean(rating))
comedy_avg_year <-
  movies %>%
  filter(Comedy == 1, year %in% 2000:2005) %>%
  summarize(comedyMeanRating = mean(rating))
documentary_avg_year <-
  movies %>%
  filter(Documentary == 1, year %in% 2000:2005) %>%
  summarize(documentaryMeanRating = mean(rating))
drama_avg_year <-
  movies %>%
  filter(Drama == 1, year %in% 2000:2005) %>%
  summarize(dramaMeanRating = mean(rating))
romance_avg_year <-
  movies %>%
  filter(Romance == 1, year %in% 2000:2005) %>%
  summarize(romanceMeanRating = mean(rating))
short_avg_year <-
  movies %>%
  filter(Short == 1, year %in% 2000:2005) %>%
  summarize(shortMeanRating = mean(rating))

action_avg_year_df <- data.frame(Genre = "Action",
                                 Rating = action_avg_year$actionMeanRating)
animation_avg_year_df <- data.frame(Genre = "Animation",
                                    Rating = animation_avg_year$animationMeanRating)
comedy_avg_year_df <- data.frame(Genre = "Comedy",
                                 Rating = comedy_avg_year$comedyMeanRating)
documentary_avg_year_df <- data.frame(Genre = "Documentary",
                                      Rating = documentary_avg_year$documentaryMeanRating)
drama_avg_year_df <- data.frame(Genre = "Drama",
                                Rating = drama_avg_year$dramaMeanRating)
romance_avg_year_df <- data.frame(Genre = "Romance",
                                  Rating = romance_avg_year$romanceMeanRating)
```

```r
short_avg_year_df <- data.frame(Genre = "Short",
                                Rating = short_avg_year$shortMeanRating)

movie_genre_rating_year <- bind_rows(action_avg_year_df,
                                animation_avg_year_df,
                                comedy_avg_year_df,
                                documentary_avg_year_df,
                                drama_avg_year_df,
                                romance_avg_year_df,
                                short_avg_year_df
                                )

ggplot(data = movie_genre_rating_year, aes(x = Genre, y = Rating, fill = Genre)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(Rating,2)), vjust = 1.5, size = 3, color = "white") +
  scale_fill_manual(values = c(
    "Action" = "#003f5c",
    "Animation" = "#374c80",
    "Comedy" = "#7a5195",
    "Documentary" = "#bc5090",
    "Drama" = "#ef5675",
    "Romance" = "#ff764a",
    "Short" = "#ffa600"
  )) +
  labs( title = "Average Ratings of Movies Produced in 2000-2005 Across Different Genres",
        x = ("Genre"), y = ("Rating"))
```



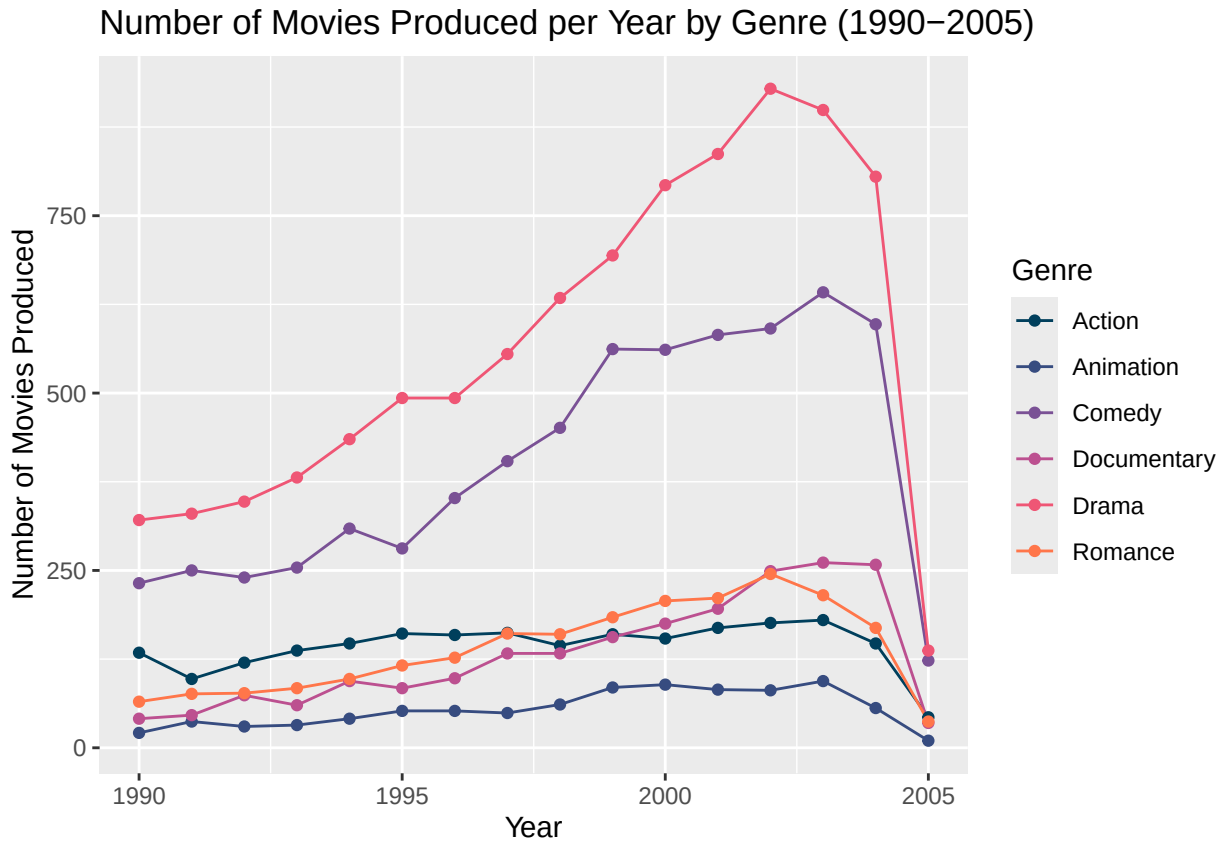
Response 7:
The average rating per genre produced in 2000-2005 is as approximately:
- Action: 5.62 - Animation: 6.56
- Comedy: 6.15
- Documentary: 7.09
- Drama: 6.37
- Romance: 6.14
- Short: 6.89

Question 8:
For each of the first 6 genres (not including short movies) consider only movies from 1990 until the last year recorded and plot a function of the number of movies in this data base of corresponding genre produced by year, for years from 1990 until the last year recorded. For each of the 6 genres you should have one curve, and plot all the curves in the same figure. Naturally, use different colors, and appropriate legend.

```
action_1990_2005 <- movies %>%
  filter(year %in% 1990:2005, Action == 1) %>%
  group_by(year) %>%
  summarize(count = n())

animation_1990_2005 <- movies %>%
  filter(year %in% 1990:2005, Animation == 1) %>%
  group_by(year) %>%
  summarize(count = n())

comedy_1990_2005 <- movies %>%
```

```
  filter(year %in% 1990:2005, Comedy == 1) %>%
  group_by(year) %>%
  summarize(count = n())

documentary_1990_2005 <- movies %>%
  filter(year %in% 1990:2005, Documentary == 1) %>%
  group_by(year) %>%
  summarize(count = n())

drama_1990_2005 <- movies %>%
  filter(year %in% 1990:2005, Drama == 1) %>%
  group_by(year) %>%
  summarize(count = n())

romance_1990_2005 <- movies %>%
  filter(year %in% 1990:2005, Romance == 1) %>%
  group_by(year) %>%
  summarize(count = n())

action_1990_2005_df <- data.frame(Genre = "Action",
                                  Count = action_1990_2005$count,
                                  Year = action_1990_2005$year)
animation_1990_2005_df <- data.frame(Genre = "Animation",
                                     Count = animation_1990_2005$count,
                                     Year = animation_1990_2005$year)
comedy_1990_2005_df <- data.frame(Genre = "Comedy",
                                  Count = comedy_1990_2005$count,
                                  Year = comedy_1990_2005$year)
documentary_1990_2005_df <- data.frame(Genre = "Documentary",
                                       Count = documentary_1990_2005$count,
                                       Year = documentary_1990_2005$year)
drama_1990_2005_df <- data.frame(Genre = "Drama",
                                 Count = drama_1990_2005$count,
                                 Year = drama_1990_2005$year)
romance_1990_2005_df <- data.frame(Genre = "Romance",
                                   Count = romance_1990_2005$count,
                                   Year = romance_1990_2005$year)

movies_1990_2005 <- bind_rows(action_1990_2005_df,
                              animation_1990_2005_df,
                              comedy_1990_2005_df,
                              documentary_1990_2005_df,
                              drama_1990_2005_df,
                              romance_1990_2005_df,
                              )

ggplot(data = movies_1990_2005, aes(x = Year, y = Count, color = Genre)) +
  geom_line() +
  geom_point() +
  scale_color_manual(values = c(
    "Action" = "#003f5c",
    "Animation" = "#374c80",
    "Comedy" = "#7a5195",
```

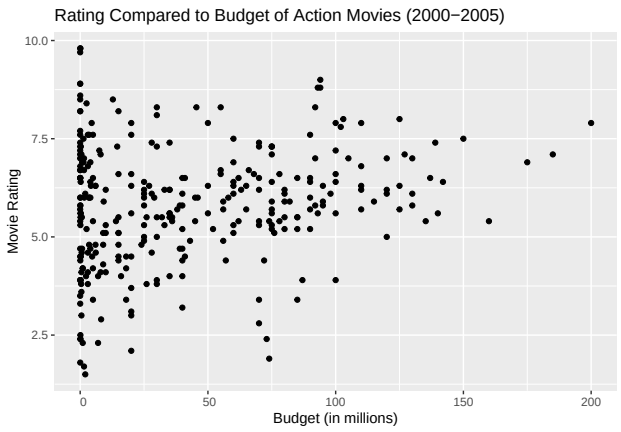
```
    "Documentary" = "#bc5090",
    "Drama" = "#ef5675",
    "Romance" = "#ff764a"
  )) +
  labs( title = "Number of Movies Produced per Year by Genre (1990-2005)",
        x = ("Year"), y = ("Number of Movies Produced"))
```

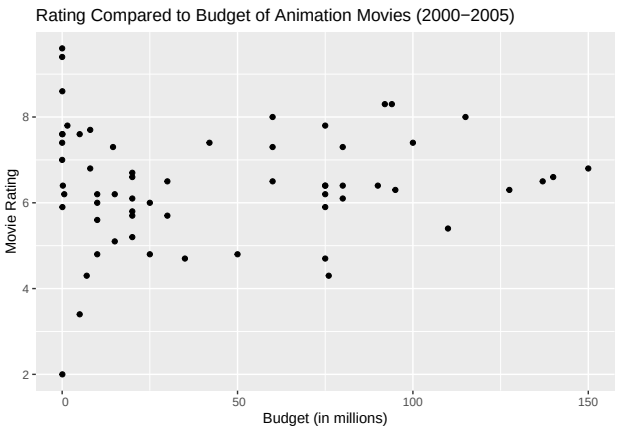


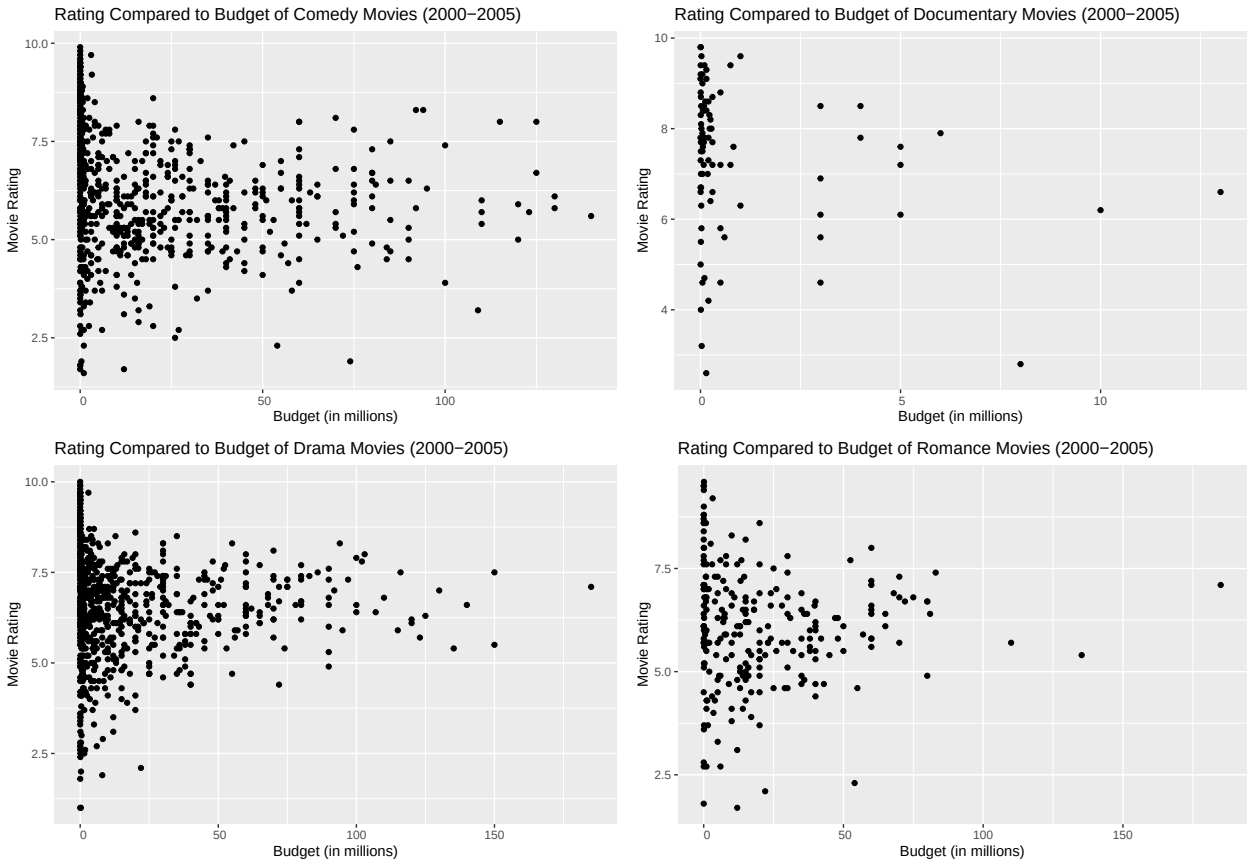
Question 9: Finally, formulate 3 questions of your choice related to this dataset and answer them.

```
a1 <- movies %>%
  filter(Action == 1 & budget != 0 & year %in% 2000:2005)

ggplot(a1, aes(x = budget, y = rating))+
  geom_point() +
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Action Movies (2000-2005)",
        x = ("Budget (in millions)"), y = ("Movie Rating"))

an1 <- movies %>%
  filter(Animation == 1 & budget != 0 & year %in% 2000:2005)

ggplot(an1, aes(x = budget, y = rating))+
  geom_point()+
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Animation Movies (2000-2005)",
```

```
        x = ("Budget (in millions)"), y = ("Movie Rating"))

c1 <- movies %>%
  filter(Comedy == 1 & budget != 0 & year %in% 2000:2005)

ggplot(c1, aes(x = budget, y = rating))+
  geom_point()+
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Comedy Movies (2000-2005)",
        x = ("Budget (in millions)"), y = ("Movie Rating"))

d1 <- movies %>%
  filter(Documentary == 1 & budget != 0 & year %in% 2000:2005)

ggplot(d1, aes(x = budget, y = rating))+
  geom_point()+
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Documentary Movies (2000-2005)",
        x = ("Budget (in millions)"), y = ("Movie Rating"))

dr1 <- movies %>%
  filter(Drama == 1 & budget != 0 & year %in% 2000:2005)

ggplot(dr1, aes(x = budget, y = rating))+
  geom_point()+
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Drama Movies (2000-2005)",
        x = ("Budget (in millions)"), y = ("Movie Rating"))

r1 <- movies %>%
  filter(Romance == 1 & budget != 0 & year %in% 2000:2005)

ggplot(r1, aes(x = budget, y = rating))+
  geom_point()+
  scale_x_continuous(labels = function(x) format(x / 1000000, scientific = FALSE))+
  labs( title = "Rating Compared to Budget of Romance Movies (2000-2005)",
        x = ("Budget (in millions)"), y = ("Movie Rating"))
```







Response 9.1:

(1) I was curious to see how different budget levels affected the ratings of movies across different genres (excluding short movies) . I limited my data scope to movies produced in the 2000s, specifically 2000 - 2005. As depicted by the scatter plots a larger budget surprisingly had very little impact on the resulting rating of the movies. Take for instance Documentaries: There is a large concentration of movies rating 8 - 10 that had a budget of less than a million, and the two movies with a budget of 10 million and greater actually scored fairly average ratings of around 6. There also tends to be greater variation in ratings with movies of lower budgets vs higher budgets. You can really see this in the action movies scatter plot. Movies with a budget of less than 10 million had a large range of ratings spanning 1.5 on the low end and 9.8 on the higher. However movie ratings in the right half of the plot had very little fluctuations with one another, usually only 2 point differences. We can conclude that a higher budget will not result in a higher rating. However a higher budget will result in consistency among ratings. Meaning that lower budget movies can fluctuate greatly in their ratings, but higher budget movies vary little in their ratings to one another.

Response 9.2:

(2) How many movies had a perfect 10 rating? What percent of total movies had a perfect score?

```
top_ratings <- movies %>% filter(rating == 10) %>% select(title, rating)
print(top_ratings)
```

```
| # A tibble: 3 x 2
|   title                                rating
|   <chr>                                 <dbl>
| 1 Dimensia Minds Trilogy: The Hope Factor     10
| 2 Fishing for Love                            10
| 3 Summer Sonata, A                            10
```

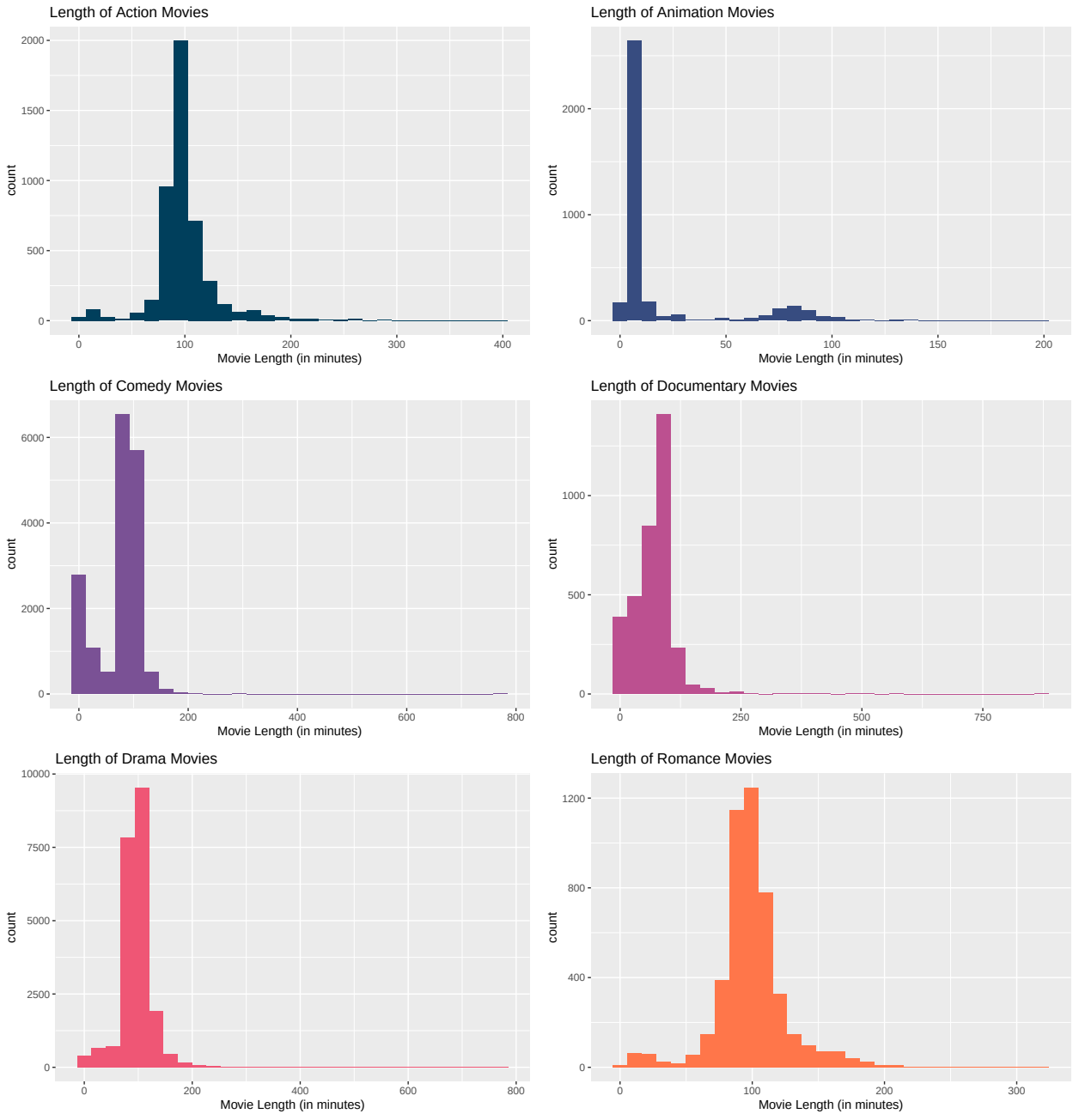
```r
top_3 <- nrow(top_ratings)
observations <- nrow(movies)

percentage <- ( top_3 / observations) * 100
formatted_percentage <- sprintf("%.2f%%", percentage)
```

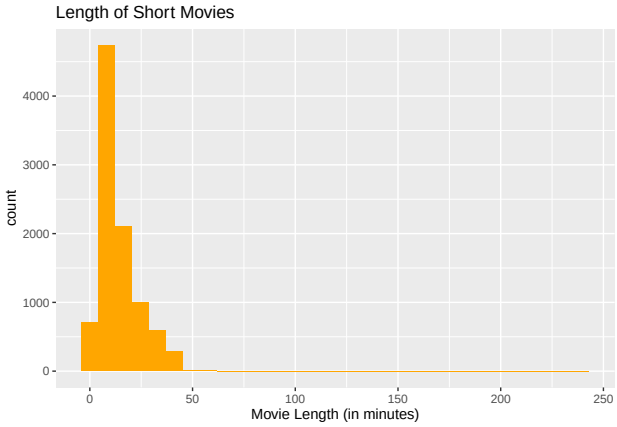
Of the total 58788 movies only 3 had a perfect rating of 10. That's approximately 0.01%.

Response 9.3:
(3) How does movie length compare to genre?

```r
a2 <- movies %>%
  filter(Action == 1)

ggplot(a2, aes(x = length, ))+
  geom_histogram(fill = "#003f5c")+
  labs( title = "Length of Action Movies",
        x = "Movie Length (in minutes)")

an2 <- movies %>%
  filter(Animation == 1)

ggplot(an2, aes(x = length, ))+
  geom_histogram(fill = "#374c80")+
  labs( title = "Length of Animation Movies",
        x = "Movie Length (in minutes)")

c2 <- movies %>%
  filter(Comedy == 1)

ggplot(c2, aes(x = length, ))+
  geom_histogram(fill = "#7a5195")+
  labs( title = "Length of Comedy Movies",
        x = "Movie Length (in minutes)")

d2 <- movies %>%
  filter(Documentary == 1)

ggplot(d2, aes(x = length, ))+
  geom_histogram(fill = "#bc5090")+
  labs( title = "Length of Documentary Movies",
        x = "Movie Length (in minutes)")

dr2 <- movies %>%
  filter(Drama == 1)

ggplot(dr2, aes(x = length, ))+
  geom_histogram(fill = "#ef5675")+
  labs( title = "Length of Drama Movies",
        x = "Movie Length (in minutes)")

r2 <- movies %>%
  filter(Romance == 1)

ggplot(r2, aes(x = length, ))+
```

```
  geom_histogram(fill = "#ff764a")+
  labs( title = "Length of Romance Movies",
        x = "Movie Length (in minutes)")


s2 <- movies %>%
  filter(Short == 1)

ggplot(s2, aes(x = length, ))+
  geom_histogram(fill = "#ffa600")+
  labs( title = "Length of Short Movies",
        x = "Movie Length (in minutes)")
```





Aside from Short movies animation movies tend to be shorter in length compared to the other genres. In contrast documentaries are longer in the length when compared to the others.