Lab2

2026-01-30

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

```
range(movies$year, na.rm = TRUE)
```

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

The movies in this dataset were produced between 1893 and 2005.
The oldest movie in the database was released in the year 1893 and the most recent movie was released in the year 2005.

```
budget_available <- sum(!is.na(movies$budget))
budget_missing <- sum(is.na(movies$budget))
total_movies <- nrow(movies)
budget_available / total_movies
```

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

```
budget_missing / total_movies
```

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

Approximately 8.87% of the movies include budget information, whereas about 91.13% of the movies do not have budget data.

```
movies %>%
  filter(!is.na(budget)) %>%
  arrange(desc(budget)) %>%
  select(title, year, budget) %>%
  head(5)
```

```
## # A tibble: 5 x 3
##   title                              year    budget
##   <chr>                             <int>     <int>
## 1 Spider-Man 2                       2004 200000000
```

```
## 2 Titanic                                1997 200000000
## 3 Troy                                   2004 185000000
## 4 Terminator 3: Rise of the Machines     2003 175000000
## 5 Waterworld                             1995 175000000
```

The five most expensive movies in the dataset (based on reported budgets) are Spider-Man 2 (2004) and Titanic (1997), with a budget of 200 million.
Next, Troy (2004) with a budget of 185 million, and Terminator 3: Rise of the Machines (2003) and Waterworld (1995), both with budgets of 175 million.

```
movies %>%
  arrange(desc(length)) %>%
  select(title, year, length) %>%
  head(5)
```

```
## # A tibble: 5 x 3
##   title                                                year length
##   <chr>                                               <int>  <int>
## 1 Cure for Insomnia, The                               1987   5220
## 2 Longest Most Meaningless Movie in the World, The     1970   2880
## 3 Four Stars                                           1967   1100
## 4 Resan                                                1987    873
## 5 Out 1                                                1971    773
```

To identify the longest movies in the dataset, the data was first sorted in descending order based on the length variable, which is the runtime of each movie in minutes. After sorting, only the movie title, year of release, and length were selected. The longest movie in the dataset is The Cure for Insomnia (1987), with a runtime of 5,220 minutes. Next, The Longest Most Meaningless Movie in the World (1970), which runs for 2,880 minutes. Other long movies are Four Stars (1967), Resan (1987), and Out 1 (1971).

```
short_movies <- movies[movies$Short == 1, ]
min(short_movies$length, na.rm = TRUE)
```

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

```
max(short_movies$length, na.rm = TRUE)
```

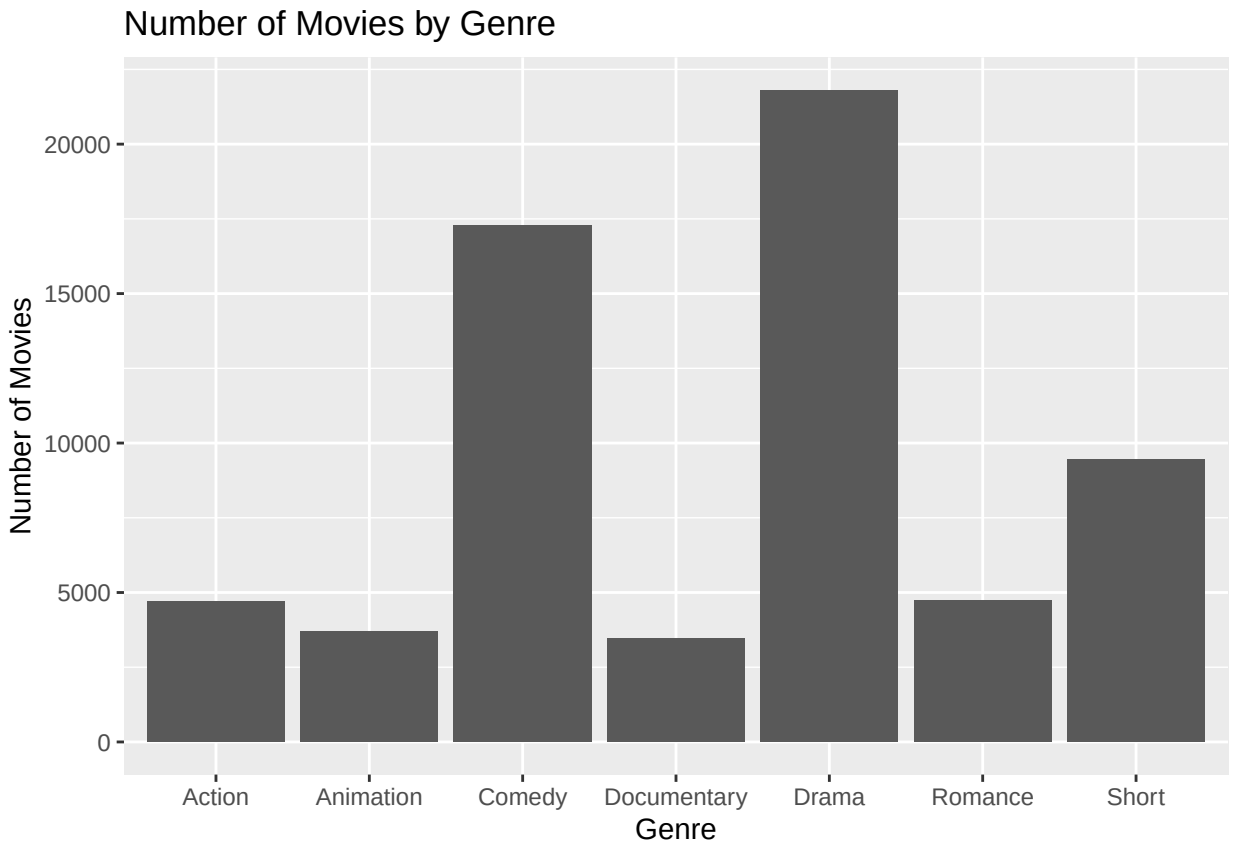
```
## [1] 240
```

```
short_movies %>%
  filter(length == min(length, na.rm = TRUE) |
         length == max(length, na.rm = TRUE)) %>%
  select(title, year, length)
```

```
## # A tibble: 166 x 3
##    title                                                  year length
##    <chr>                                                 <int>  <int>
##  1 10 jaar leuven kort                                    2004    240
##  2 17 Seconds to Sophie                                   1998      1
##  3 2 A.M. in the Subway                                   1905      1
##  4 Admiral Cigarette                                      1897      1
##  5 Admiral Dewey Leading Land Parade                      1899      1
##  6 Alphonse and Gaston, No. 3                             1903      1
##  7 Ameta                                                  1903      1
##  8 Amy Muller                                             1896      1
##  9 Arabian Gun Twirler                                    1899      1
## 10 Arrival of McKinley's Funeral Train at Canton, Ohio    1901      1
## # i 156 more rows
```

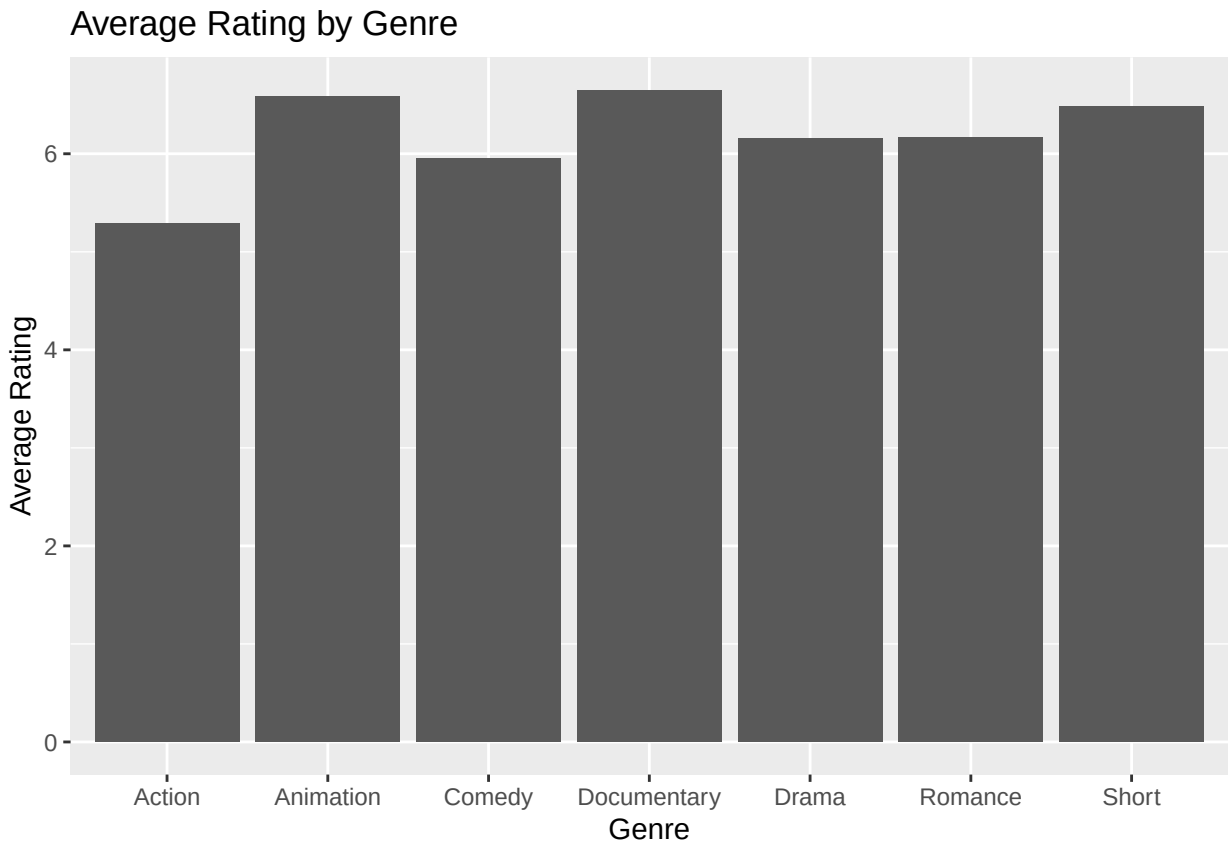
First, the dataset was subset to include only movies classified as short films by selecting rows where the variable Short equals 1. This created a new data containing only short movies. Next, the minimum and maximum values of the length variable were calculated. Finally, the dataset of short movies was filtered to display the titles, years, and lengths of the movies corresponding to these runtimes. Shortest short movie = 1 minute = 165 movies for example Wee Wee(2000) Longest short movie = 240 minutes = 10 jaar leuven kort(2004)

```
genre_counts <- movies %>%
  summarise(
    Action = sum(Action),
    Animation = sum(Animation),
    Comedy = sum(Comedy),
    Drama = sum(Drama),
    Documentary = sum(Documentary),
    Romance = sum(Romance),
    Short = sum(Short)
  ) %>%
  pivot_longer(
    everything(),
    names_to = "genre",
    values_to = "count"
  )
ggplot(genre_counts, aes(x = genre, y = count)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Number of Movies by Genre",
    x = "Genre",
    y = "Number of Movies"
  )
```



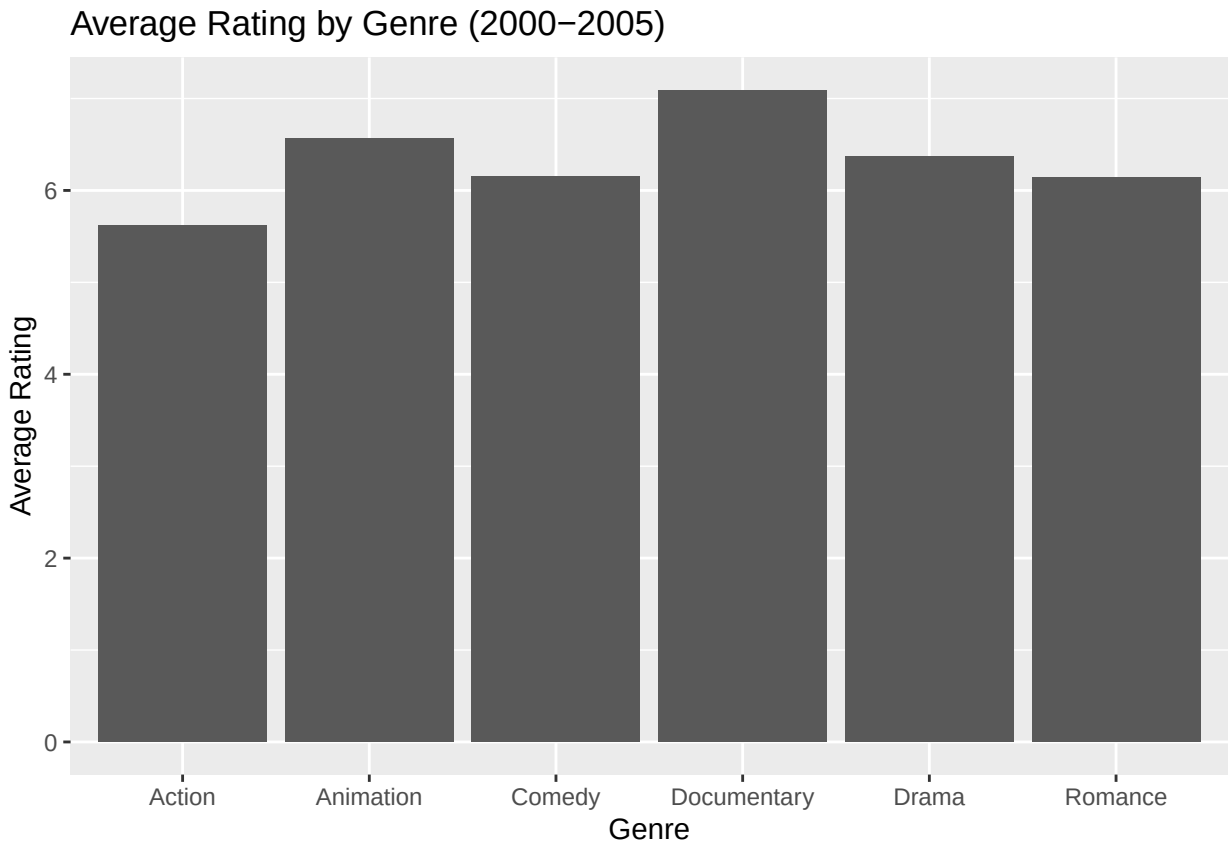
To calculate the number of movies in each genre, the dataset was summarized by adding the binary genre indicator variables. Because each genre variable is coded as 1 when a movie belongs to that genre and 0 otherwise, summing each variable gives the total number of movies in that genre. The data was then reshaped into a long format using pivot_longer() to make it suitable for plotting. Finally, a bar plot was created to compare the number of movies across genres. The bar plot shows that Drama is the most common genre in the dataset, followed by Comedy. Short films also represent a substantial part of the dataset. Whereas, Animation and Documentary films appear less frequently.

```
avg_rating_genre <- movies %>%
  summarise(
    Action = mean(rating[Action == 1], na.rm = TRUE),
    Animation = mean(rating[Animation == 1], na.rm = TRUE),
    Comedy = mean(rating[Comedy == 1], na.rm = TRUE),
    Drama = mean(rating[Drama == 1], na.rm = TRUE),
    Documentary = mean(rating[Documentary == 1], na.rm = TRUE),
    Romance = mean(rating[Romance == 1], na.rm = TRUE),
    Short = mean(rating[Short == 1], na.rm = TRUE)
  ) %>%
  pivot_longer(everything(), names_to = "genre", values_to = "avg_rating")

ggplot(avg_rating_genre, aes(x = genre, y = avg_rating)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Rating by Genre",
       x = "Genre",
       y = "Average Rating")
```



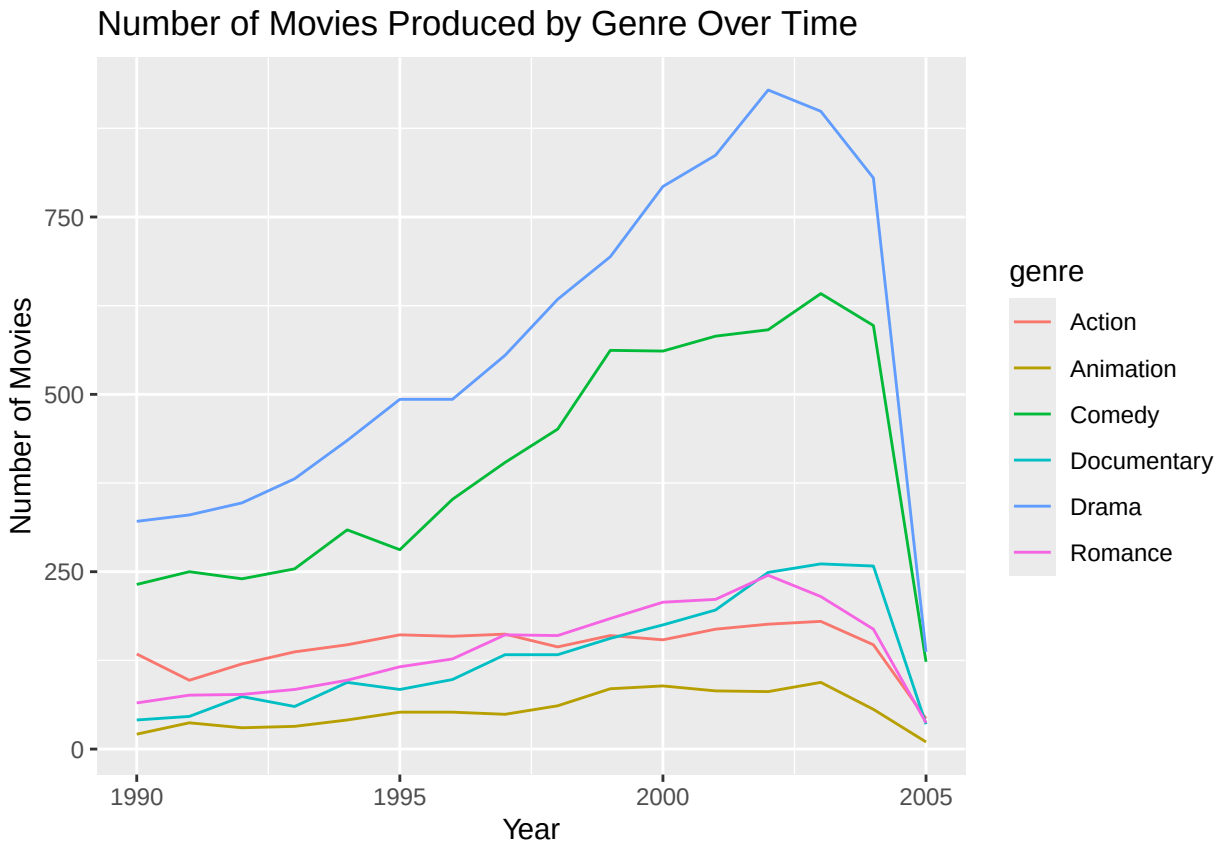
To calculate the average IMDb rating for each genre, the dataset was summarized by calculating the mean of the rating variable for movies belonging to each genre. Because genre indicators are coded as 1 when a movie belongs to a given genre, the ratings were subset using conditions such as Action == 1. Missing rating values were excluded from the calculations using na.rm = TRUE. The resulting summary was reshaped using pivot_longer(). A bar plot was then created to compare the average ratings across genres. The bar plot shows that Documentary and Animation movies have the highest average ratings. Short films also receive high average ratings compared to other genres. Whereas, Action movies have the lowest average rating.

```
movies_2000_2005 <- movies %>% filter(year >= 2000, year <= 2005)
avg_rating_2000_2005 <- movies_2000_2005 %>%
  summarise(
    Action = mean(rating[Action == 1], na.rm = TRUE),
    Animation = mean(rating[Animation == 1], na.rm = TRUE),
    Comedy = mean(rating[Comedy == 1], na.rm = TRUE),
    Drama = mean(rating[Drama == 1], na.rm = TRUE),
    Documentary = mean(rating[Documentary == 1], na.rm = TRUE),
    Romance = mean(rating[Romance == 1], na.rm = TRUE)
  ) %>%
  pivot_longer(everything(), names_to = "genre", values_to = "avg_rating")

ggplot(avg_rating_2000_2005, aes(x = genre, y = avg_rating)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Rating by Genre (2000-2005)",
       x = "Genre",
       y = "Average Rating")
```



To calculate average movie ratings by genre for a specific time period, the dataset was filtered to include only movies released between the years 2000 and 2005. For this subset of movies, the average IMDb rating was calculated separately for each genre by computing the mean of the rating variable for movies where the corresponding genre indicator equals 1. Missing rating values were excluded using na.rm = TRUE. The summarized results were then reshaped and visualized with a bar plot. The bar plot indicates that Documentary films released between 2000 and 2005 have the highest average ratings, followed by Animation and Drama. Action movies have the lowest average ratings among the genres, even within this more recent time period.

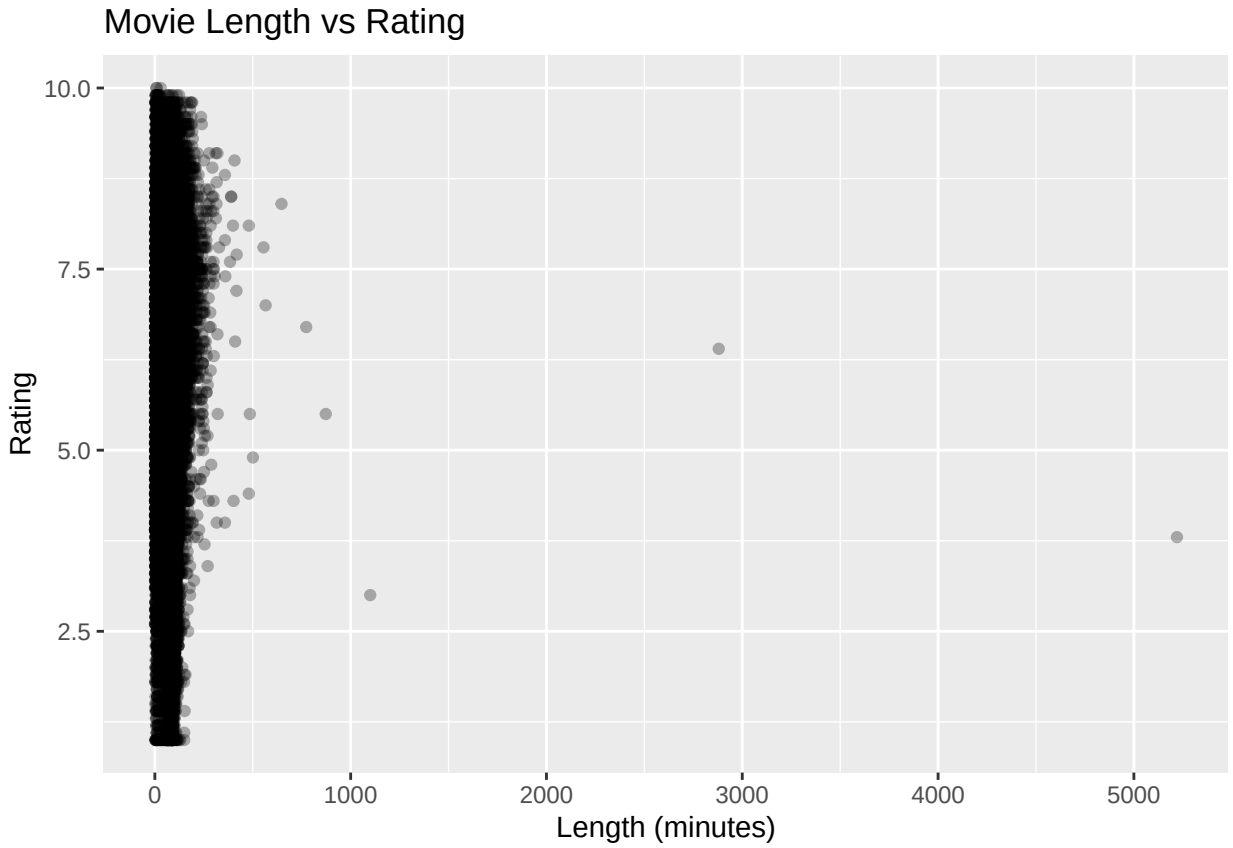
```
movies_1990 <- movies %>% filter(year >= 1990)
genre_time <- movies_1990 %>%
  select(year, Action, Animation, Comedy, Drama, Documentary, Romance) %>%
  pivot_longer(-year, names_to = "genre", values_to = "value") %>%
  filter(value == 1) %>%
  group_by(year, genre) %>%
  summarise(count = n(), .groups = "drop")
ggplot(genre_time, aes(x = year, y = count, color = genre)) +
  geom_line() +
  labs(title = "Number of Movies Produced by Genre Over Time",
       x = "Year",
       y = "Number of Movies")
```



To analyze trends in movie production by genre over time, the dataset was first formatted to movies released from 1990 onward. The relevant genre indicator variables (Action, Animation, Comedy, Drama, Documentary, and Romance) were selected along with the release year. The data was then reshaped, allowing each row to represent a single genre for a given movie. Movies belonging to a genre were identified by filtering for rows where the genre indicator equals 1. The data was grouped by year and genre, and the number of movies produced each year within each genre was counted. Finally, a multi-line plot was created. The line plot shows that Drama and Comedy movies experienced the largest growth in production from 1990 through the early 2000s, with Drama being the most frequently produced genre throughout this period. Documentary and Romance films also show steady increases over time. Animation movies remain the least common genre.

9). Question 1: Is there a relationship between movie length and IMDb rating? The scatter plot shows no strong linear relationship between movie length and rating. This suggests that longer movies do not necessarily receive higher ratings, and runtime alone is not a strong predictor of viewer satisfaction.

```
ggplot(movies, aes(x = length, y = rating)) +
  geom_point(alpha = 0.3) +
  labs(title = "Movie Length vs Rating",
       x = "Length (minutes)",
       y = "Rating")
```



Question 2: Which movie genres receive the highest average number of user votes? The results show that Action and Animation movies tend to receive the highest average number of votes. This suggests that these genres attract broader audiences and generate higher viewer participation on IMDb. In contrast, genres such as Documentary and Romance receive fewer votes on average, likely reflecting more niche audiences.

```
movies %>%
  summarise(
    Action = mean(votes[Action == 1], na.rm = TRUE),
    Animation = mean(votes[Animation == 1], na.rm = TRUE),
    Comedy = mean(votes[Comedy == 1], na.rm = TRUE),
    Drama = mean(votes[Drama == 1], na.rm = TRUE),
    Documentary = mean(votes[Documentary == 1], na.rm = TRUE),
    Romance = mean(votes[Romance == 1], na.rm = TRUE)
  ) %>%
  pivot_longer(everything(), names_to = "genre", values_to = "avg_votes")
```

```
## # A tibble: 6 x 2
##   genre       avg_votes
##   <chr>           <dbl>
## 1 Action          2087.
## 2 Animation        351.
## 3 Comedy           752.
## 4 Drama            880.
## 5 Documentary      103.
## 6 Romance         1281.
```

Question 3: How has the number of movies produced per year changed over time? The line plot shows a clear upward trend in the number of movies produced per year, particularly after the 1990s. This indicates

substantial growth in film production over time, likely driven by advances in filmmaking technology and increased accessibility to distribution platforms. The decline observed near the final year of the dataset likely reflects incomplete data.

```
movies %>%
  group_by(year) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = year, y = count)) +
  geom_line() +
  labs(title = "Number of Movies Produced Per Year",
       x = "Year",
       y = "Number of Movies")
```