IMDB Movies Data Analysis

Savaira Imran

3/06/2025

Load Required Libraries

```
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)
library(tidyr)
```

Load Data

```
data(movies)
```

1. Range of Years of Production

```
year_range <- range(movies$year)
year_range
```

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

2. Proportion of Movies with Budget Information

```
budget_proportion <- mean(!is.na(movies$budget))
no_budget_proportion <- mean(is.na(movies$budget))

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

budget_proportion
```

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

```
no_budget_proportion
```

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

```
expensive_movies
```

```
## # 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
```

3. Top 5 Longest Movies

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

```
## # 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
```

4. Shortest and Longest Short Movies

```
short_movies <- movies %>% filter(if_any(starts_with("Short"), ~ . == 1))
shortest_short <- short_movies %>% arrange(length) %>% select(title, length) %>% head(1)
longest_short <- short_movies %>% arrange(desc(length)) %>% select(title, length) %>% head(1)

shortest_short
```

```
## # A tibble: 1 x 2
##   title                length
##   <chr>                 <int>
## 1 17 Seconds to Sophie      1
```

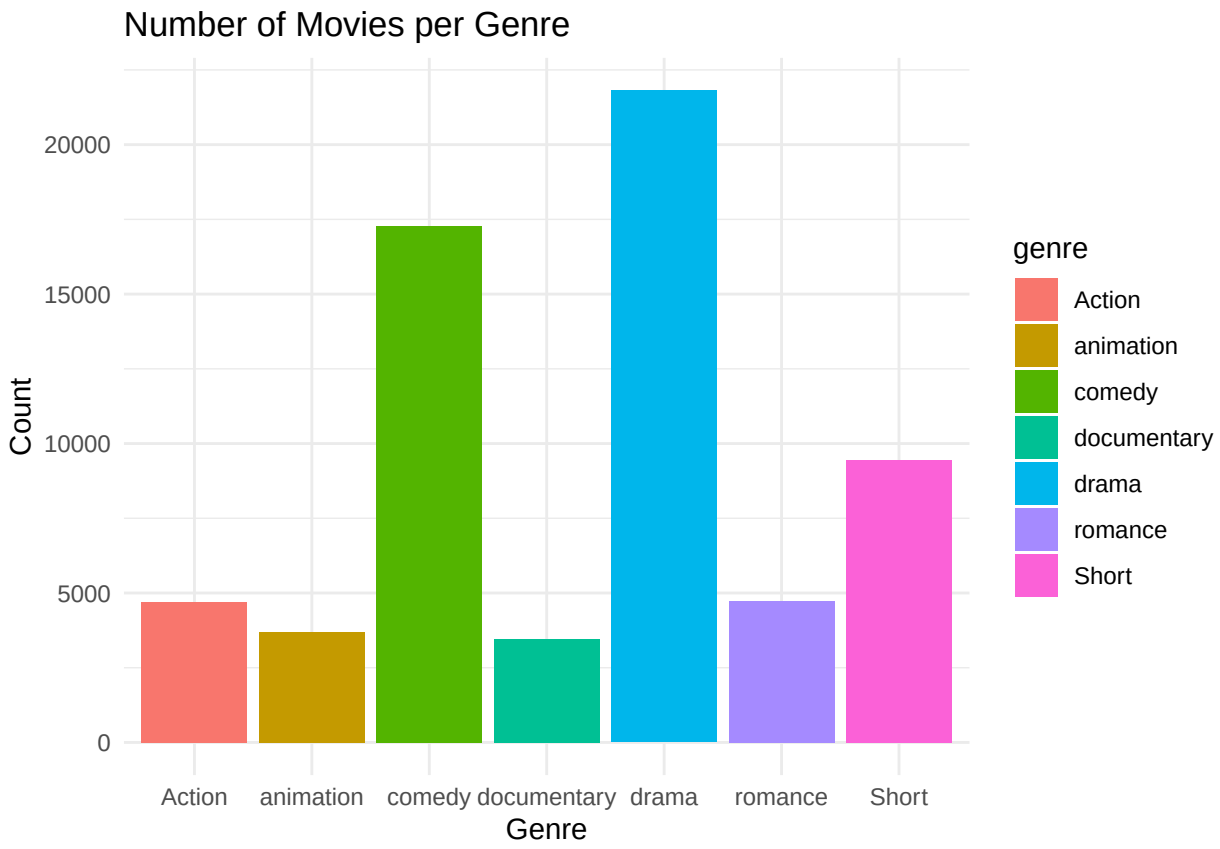
```
longest_short
```

```
## # A tibble: 1 x 2
##   title               length
##   <chr>                <int>
## 1 10 jaar leuven kort    240
```

5. Number of Movies per Genre

```
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(cols = everything(), names_to = "genre", values_to = "count")

ggplot(genre_counts, aes(x = genre, y = count, fill = genre)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Number of Movies per Genre", x = "Genre", y = "Count")
```

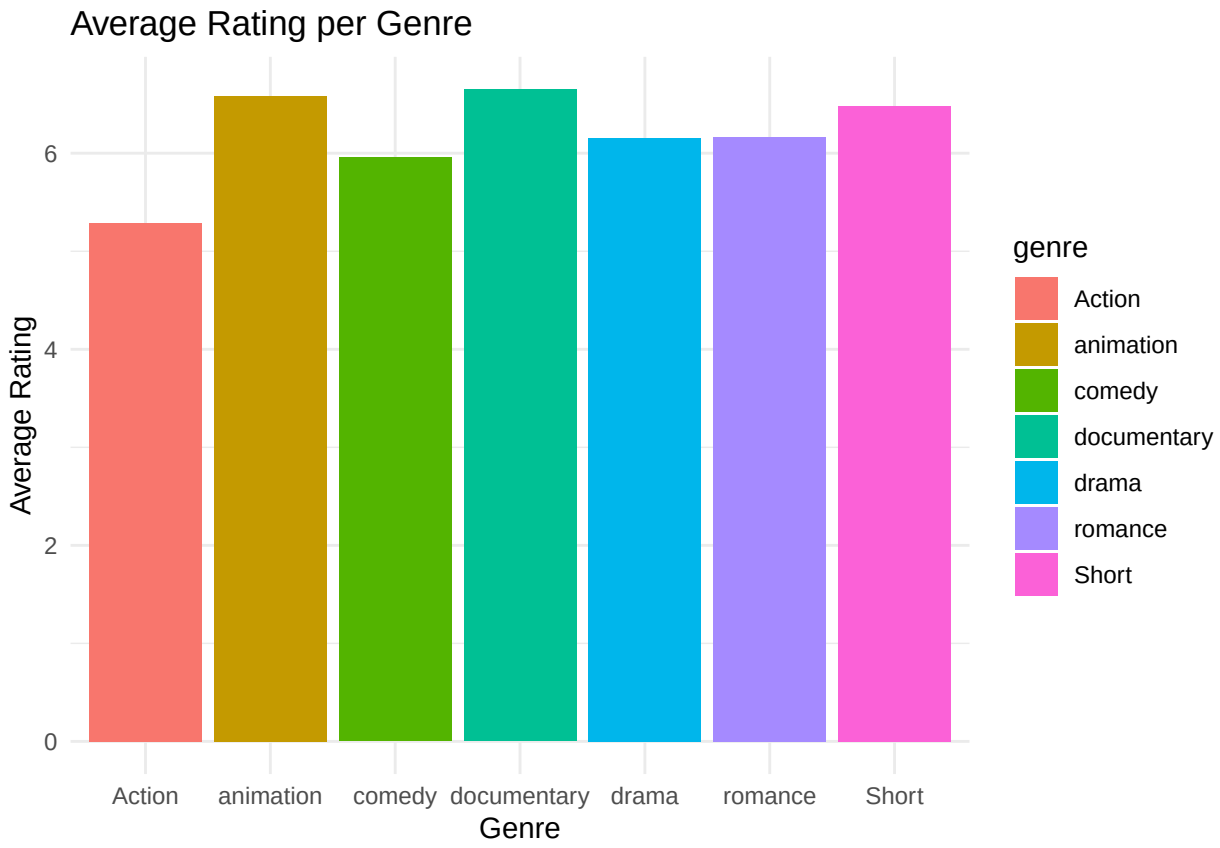


6. Average Rating by Genre

```
ratings_per_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(cols = everything(), names_to = "genre", values_to = "avg_rating")

ggplot(ratings_per_genre, aes(x = genre, y = avg_rating, fill = genre)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average Rating per Genre", x = "Genre", y = "Average Rating")
```

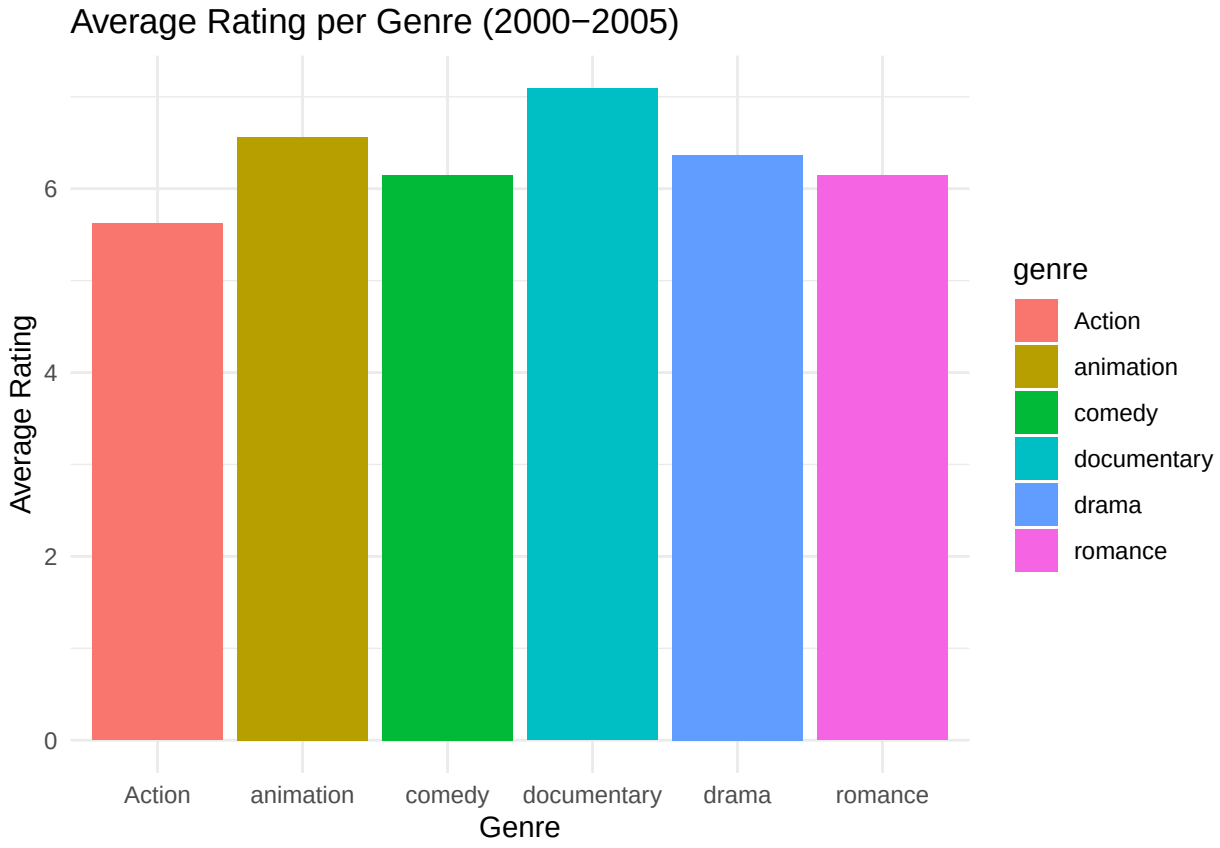


7. Average Rating of Movies (2000-2005)

```
avg_rating_2000_2005 <- movies %>%
  filter(year >= 2000 & year <= 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(cols = everything(), names_to = "genre", values_to = "avg_rating")

ggplot(avg_rating_2000_2005, aes(x = genre, y = avg_rating, fill = genre)) +
```

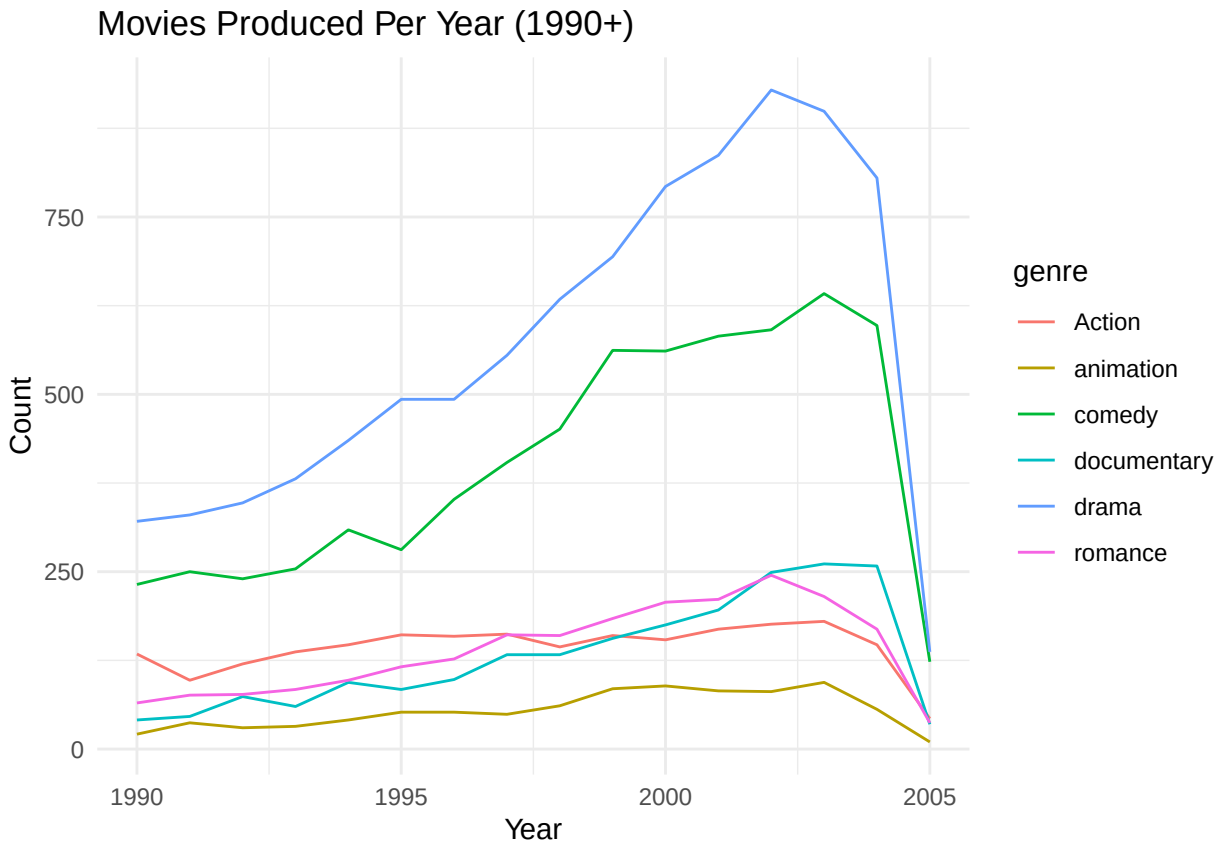
```
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Average Rating per Genre (2000-2005)", x = "Genre", y = "Average Rating")
```



8. Movie Production Trends (1990+)

```
movies_1990 <- movies %>%
  filter(year >= 1990) %>%
  group_by(year) %>%
  summarise(
    Action = sum(Action),
    animation = sum(Animation),
    comedy = sum(Comedy),
    drama = sum(Drama),
    documentary = sum(Documentary),
    romance = sum(Romance)
  ) %>%
  pivot_longer(cols = -year, names_to = "genre", values_to = "count")

ggplot(movies_1990, aes(x = year, y = count, color = genre)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Movies Produced Per Year (1990+)", x = "Year", y = "Count")
```



9. Custom Questions

a) Which genre has the highest-rated movie?

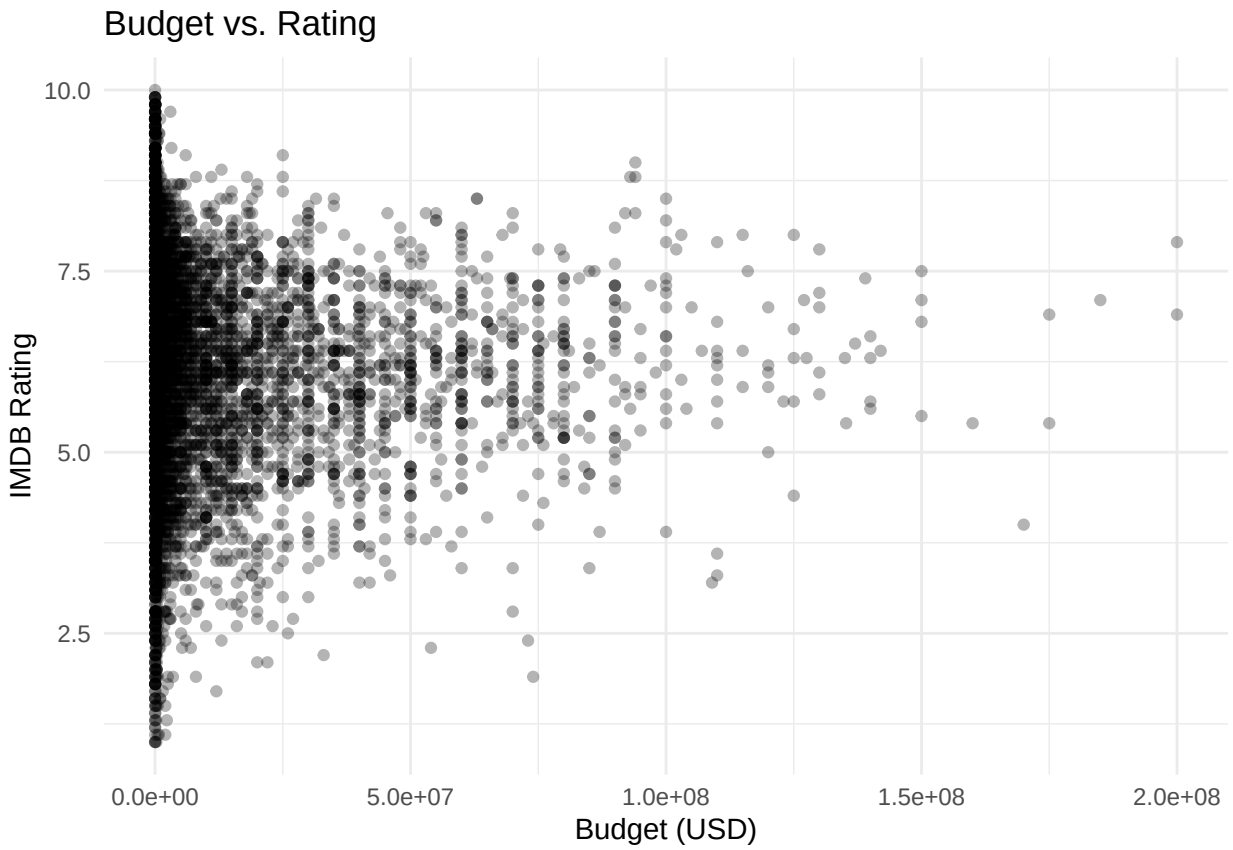
```
highest_rated <- movies %>%
  arrange(desc(rating)) %>%
  select(title, rating) %>%
  head(1)
highest_rated
```

```
## # A tibble: 1 x 2
##   title                                    rating
##   <chr>                                     <dbl>
## 1 Dimensia Minds Trilogy: The Hope Factor     10
```

b) How does budget correlate with ratings?

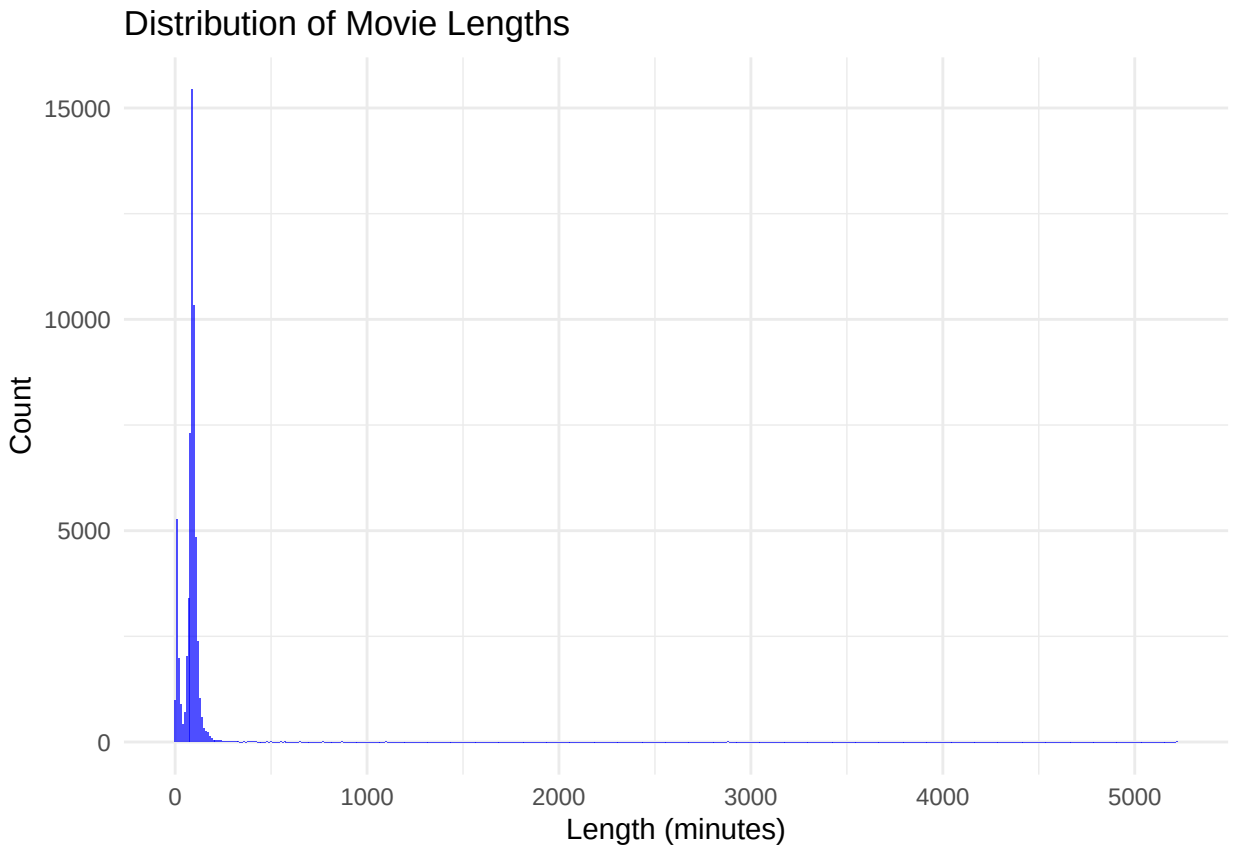
```
ggplot(movies, aes(x = budget, y = rating)) +
  geom_point(alpha = 0.3) +
  theme_minimal() +
  labs(title = "Budget vs. Rating", x = "Budget (USD)", y = "IMDB Rating")
```

```
## Warning: Removed 53573 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



c) What is the distribution of movie lengths?

```
ggplot(movies, aes(x = length)) +
  geom_histogram(binwidth = 10, fill = "blue", alpha = 0.7) +
  theme_minimal() +
  labs(title = "Distribution of Movie Lengths", x = "Length (minutes)", y = "Count")
```



This report provides insights into the IMDB movies dataset with visualizations and statistical summaries.