Analysis_IMDB_MovieDataset

Sam Kazan

2023-01-20

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE, comment = NA)
```

```
library(ggplot2movies)
library(dplyr)
library(ggplot2)
data(movies)
##movies
```

Question 1

```
range_years <- range(movies %>% select(year) %>% pull())
```

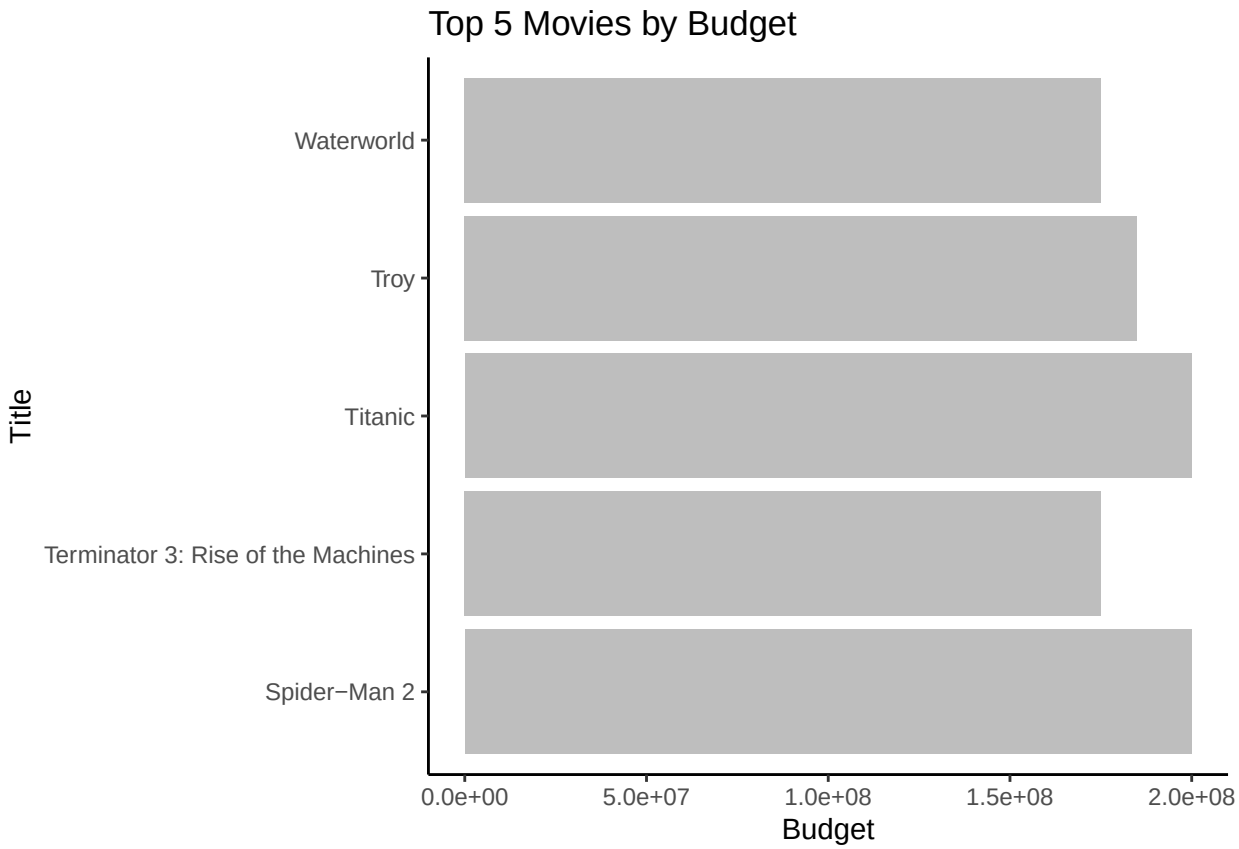
Movies in the database range from the year 1893 to year 2005

Question 2

```
budget_NA <- movies %>% filter(is.na(budget)) %>% nrow()
budget_no_NA <- movies %>% filter(!is.na(budget)) %>% nrow()
budgetcount <- c(budget_NA,budget_no_NA)
total_movie <- movies %>%  nrow()
```

There are 53573 movies that does not have a defined total budget while 5215 movies have listed total budget number. 8.870858 % of the movies don't have a budget included in the dataset. On the other hand 91.129142 % of the movies have a budget included.

```
dftop5 <- movies %>%
filter(!is.na(budget)) %>%
arrange(desc(budget)) %>%
  top_n(5, budget) %>%
  select(title, budget)

ggplot(data=dftop5, aes(x = title, y = budget)) +
  geom_bar(stat = "identity", fill = "grey") +
  xlab("Title") + ylab("Budget") + ggtitle("Top 5 Movies by Budget") +
  theme_classic() +
  coord_flip()
```



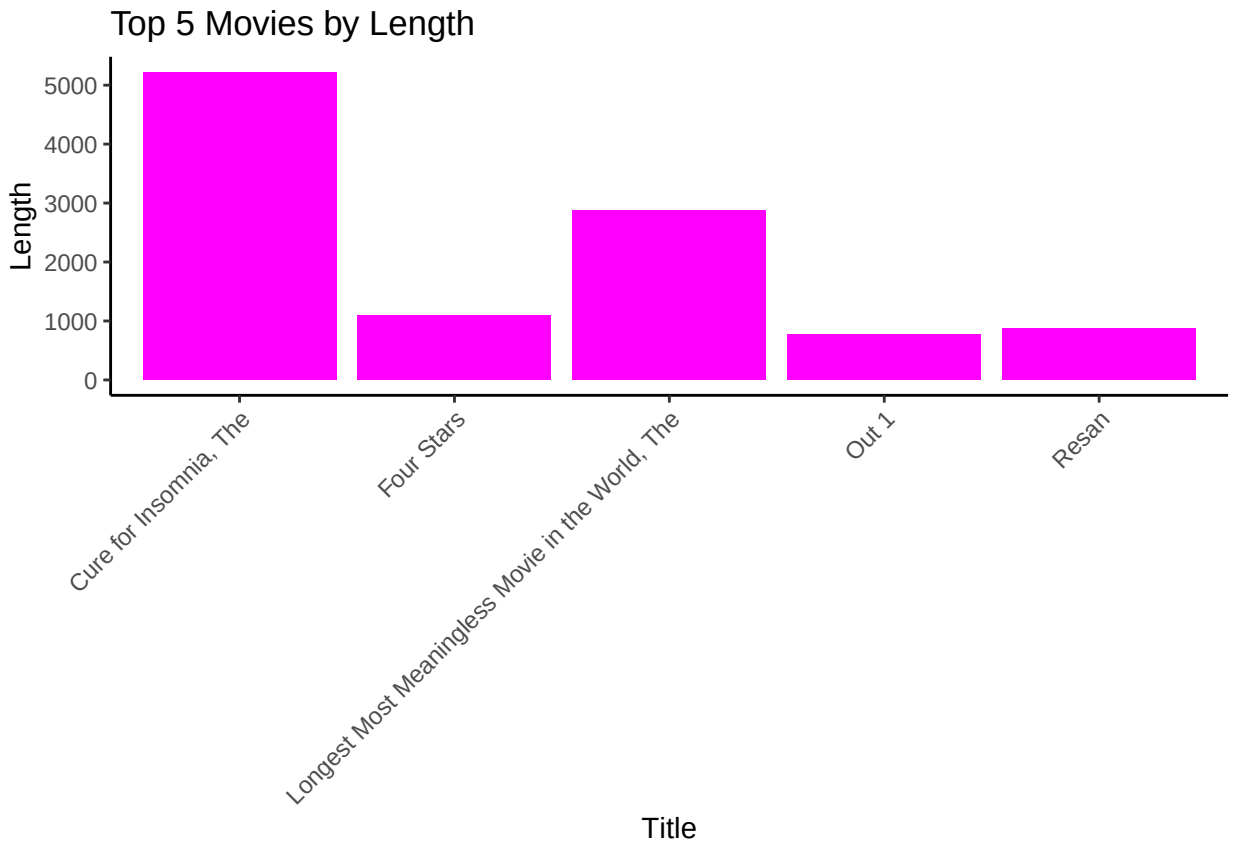

```
TopMoviesByBudget <- dftop5 %>% pull(title)
```

Top 5 most expensive movies in this data set are Spider-Man 2, Titanic, Troy, Terminator 3: Rise of the Machines, Waterworld

Question 3 & 4

```
dflongest <-  movies %>%
              select(title, length) %>%
              arrange(desc(length)) %>%
              top_n(5, length)
LongestMovies <- dflongest %>% pull(title)

dfshorts <- movies %>%
            select(title, length, Short) %>%
            filter(Short == 1) %>%
            arrange(desc(length))


shortmovies <- dfshorts %>%
  select(length) %>%
  filter(length == 1) %>%
  summarize(count = n())

ggplot(data=dflongest, aes(x = title, y = length)) +
  geom_bar(stat = "identity", fill = "magenta") +
  xlab("Title") + ylab("Length") + ggtitle("Top 5 Movies by Length") +
```

```
  theme_classic()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



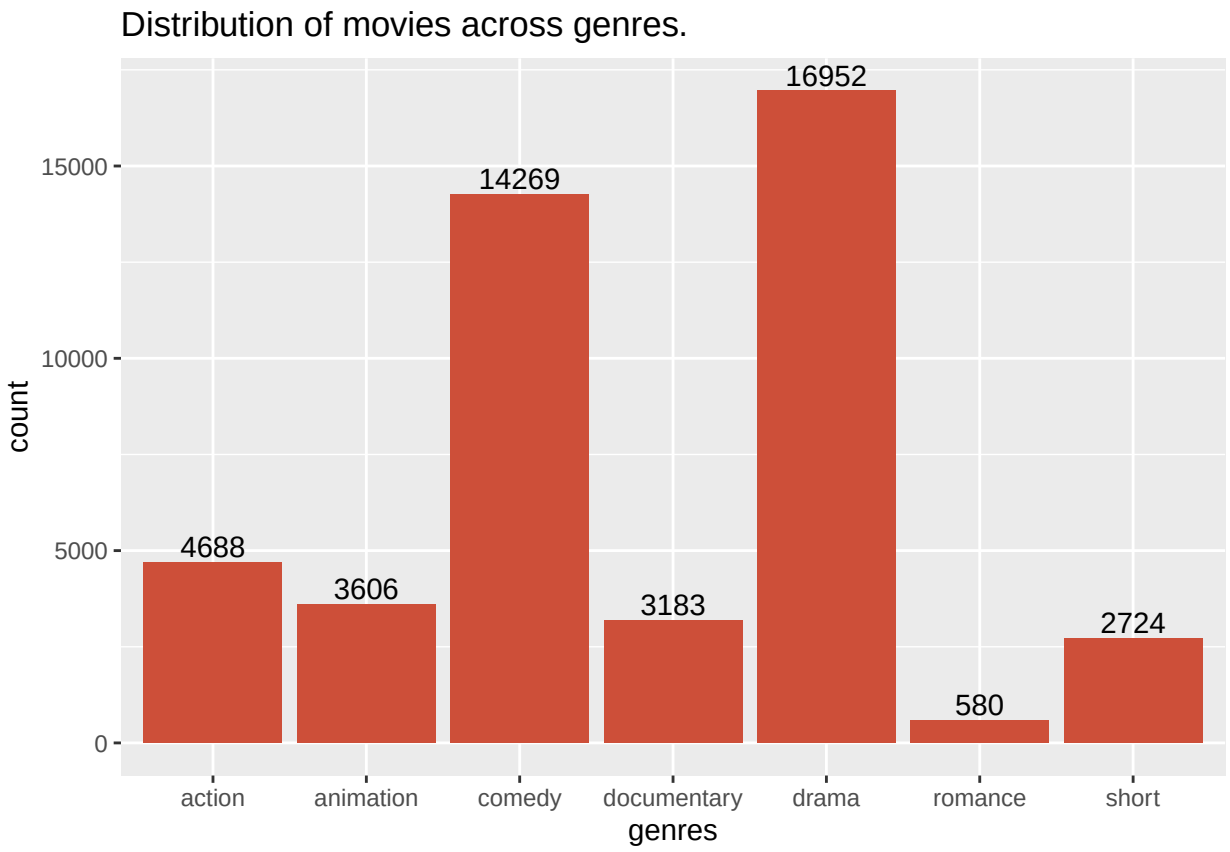
Top 5 longest movies are Cure for Insomnia, The, Longest Most Meaningless Movie in the World, The, Four Stars, Resan, Out 1. The longest movies are Cure for Insomnia, The and Four Stars which triple the length of the other 3 movies that are in the top 5 list.

Looking at all the short movies at the list the longest short movie is 10 jaar leuven kort with the running time 240 mins. There are 165 short movies that have the running time of 1 mins. Those movies are the shortest short movies in the database.

Question 5

```
dfGenres <- movies %>%
  mutate(genres = case_when(
    Action == 1 ~ "action",
    Animation == 1 ~ "animation",
    Comedy == 1 ~ "comedy",
    Drama == 1 ~ "drama",
    Documentary == 1 ~ "documentary",
    Romance == 1 ~ "romance",
    Short == 1 ~ "short"
  )) %>%
  filter(!is.na(genres))

dfGenres %>%
```

```
  group_by(genres) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = genres, y = count)) +
  ggtitle("Distribution of movies across genres.")+
  geom_col(fill = "tomato3") +
  geom_text(aes(label = count), position = position_dodge(),vjust = -0.25, color = "black")
```



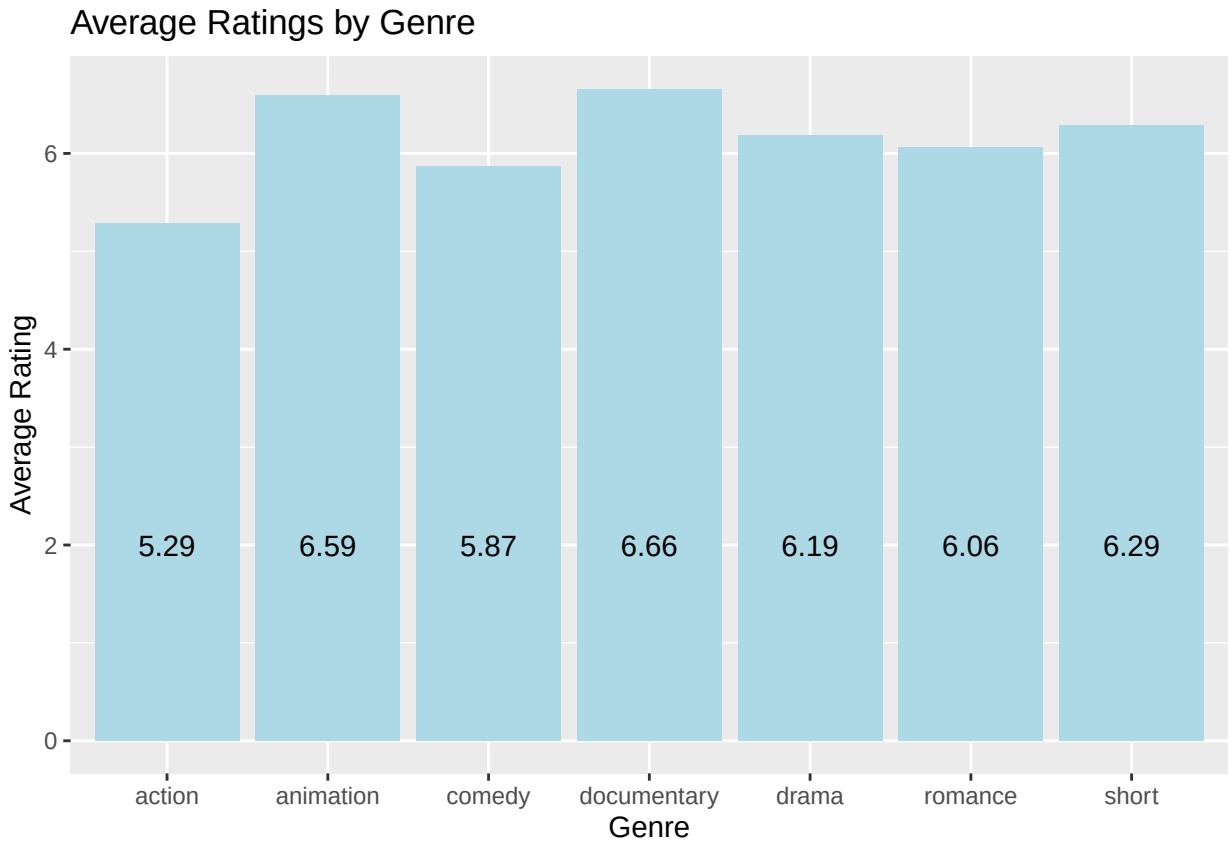

```
genre_coutn <- dfGenres %>%
  group_by(genres) %>%
  count() %>%
  pull()

genres_list = dfGenres %>% distinct(genres) %>% arrange(genres) %>% pull(genres)
```

There are 7 classification of movies: action, animation, comedy, documentary, drama, romance, short. NA are the ones without a specific genre listed in the database and are not included int the plot. Comedy and Drama are the most produced genres with movie counts of 14269 and 16952, respectively.

Question 6

```
dfGenres_avg <- dfGenres %>%
                group_by(genres) %>%
                summarize(avg_rating = round(mean(rating), 2))
##dfGenres_avg
average_rating = dfGenres_avg %>% pull(avg_rating)

ggplot(data = dfGenres_avg, aes(x = genres, y = avg_rating)) +
```

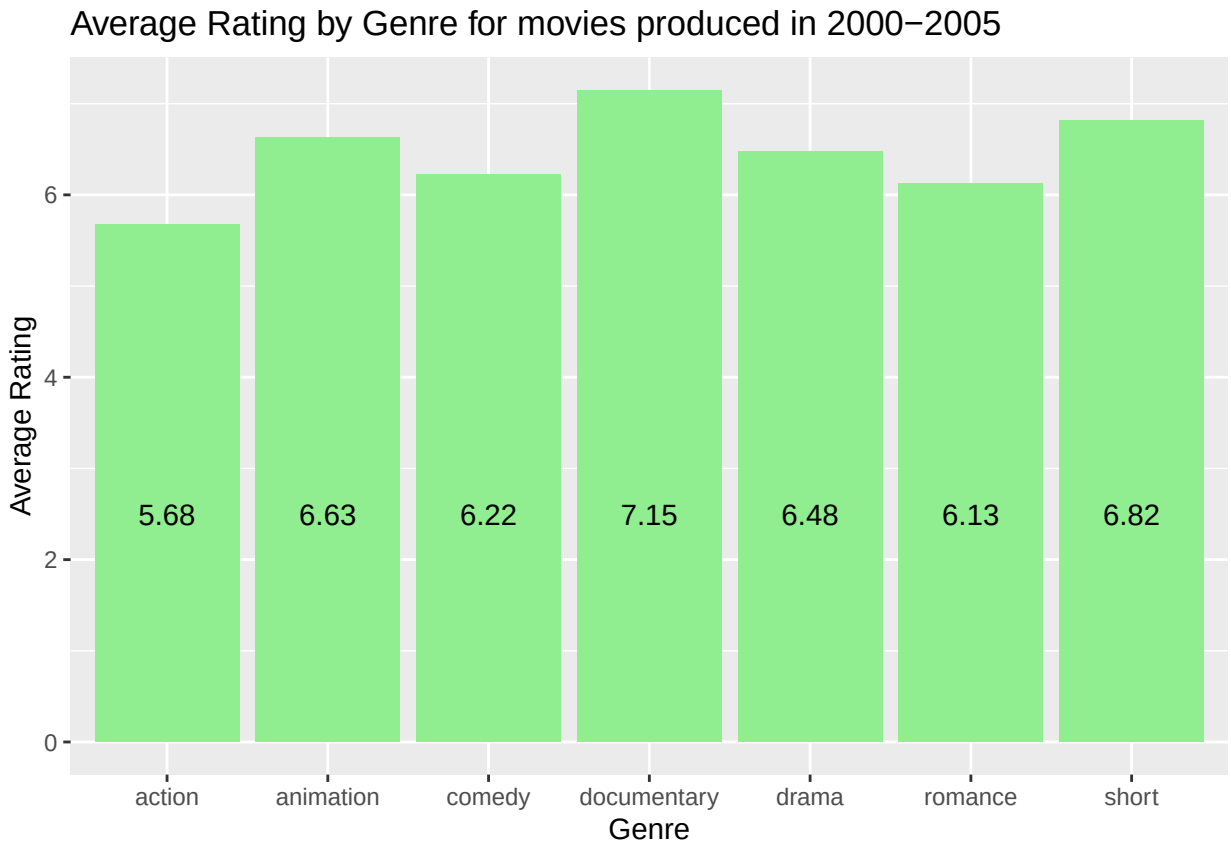
```
  geom_bar(stat = "identity") +
  ggtitle("Average Ratings by Genre")+
  xlab("Genre") +
  ylab("Average Rating") +
  geom_col(fill = "lightblue") +
  geom_text(aes(label = avg_rating), position = position_fill(vjust = 2), color = "black") +
  scale_linetype_manual(name = "Average", values = c("Average" = "dashed"))
```



All of the genres have average ratings between 5.29 and 6.66. Documentaries and Dramas are most liked movie genres with average ratings 6.66 and 6.19 consecutively.

Question 7

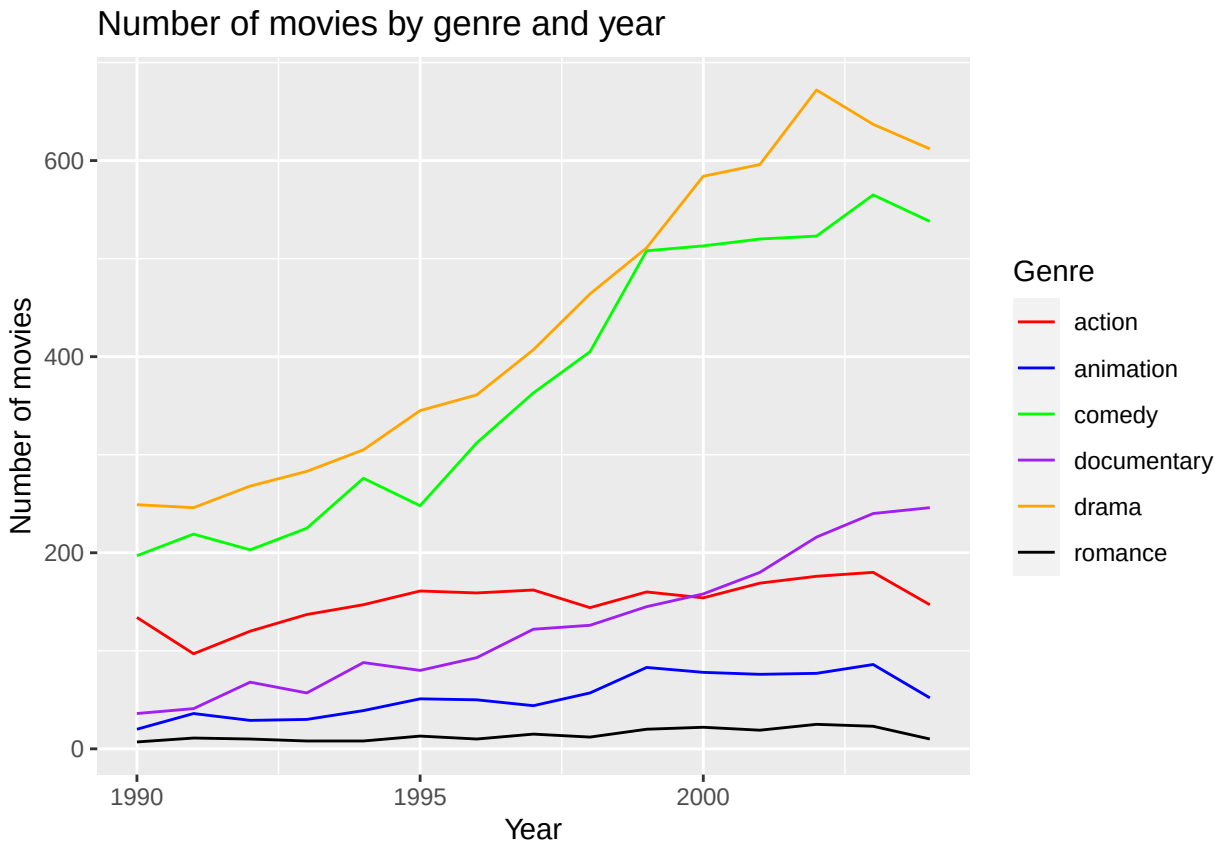
```
df_genres2000 <- dfGenres %>%
  filter( year >+ 2000, year<2006) %>%
  group_by(genres) %>%
  summarize(avg_rating_2000 = round(mean(rating),2))


ggplot(data=df_genres2000, aes(x = genres, y = avg_rating_2000)) +
  geom_col(fill= "lightgreen") +
  geom_text(aes(label = avg_rating_2000 ), position = position_fill(, vjust = 2.5)) +
  xlab("Genre") +
  ylab("Average Rating") +
  ggtitle("Average Rating by Genre for movies produced in 2000-2005")
```



Looking at the movies produced between 2000-2005 on average the genres that are most liked were documentary and short movies. Comparing this graph with all the averages by genres, we can see that short movies started getting ranked higher which surpass drama from previous dataset.

Question 8

```
movies_1990_onwards <- dfGenres %>%
                        filter(year >= 1990,year < 2005,genres != "NA",
                               genres != "short") %>%
                        group_by(genres, year) %>%
                        count()
##dfGenres
##movies_1990_onwards

ggplot(data = movies_1990_onwards, aes(x = year, y = n, color = genres)) +
  geom_line() +
  scale_color_manual(values = c("red", "blue", "green", "purple",
                                "orange", "black")) +
  labs(title = "Number of movies by genre and year", x = "Year",
       y = "Number of movies", color = "Genre")
```



The graph shows that the number of movies released in genres action and comedyare always higher (at least double) than the movies rleased under other genres.

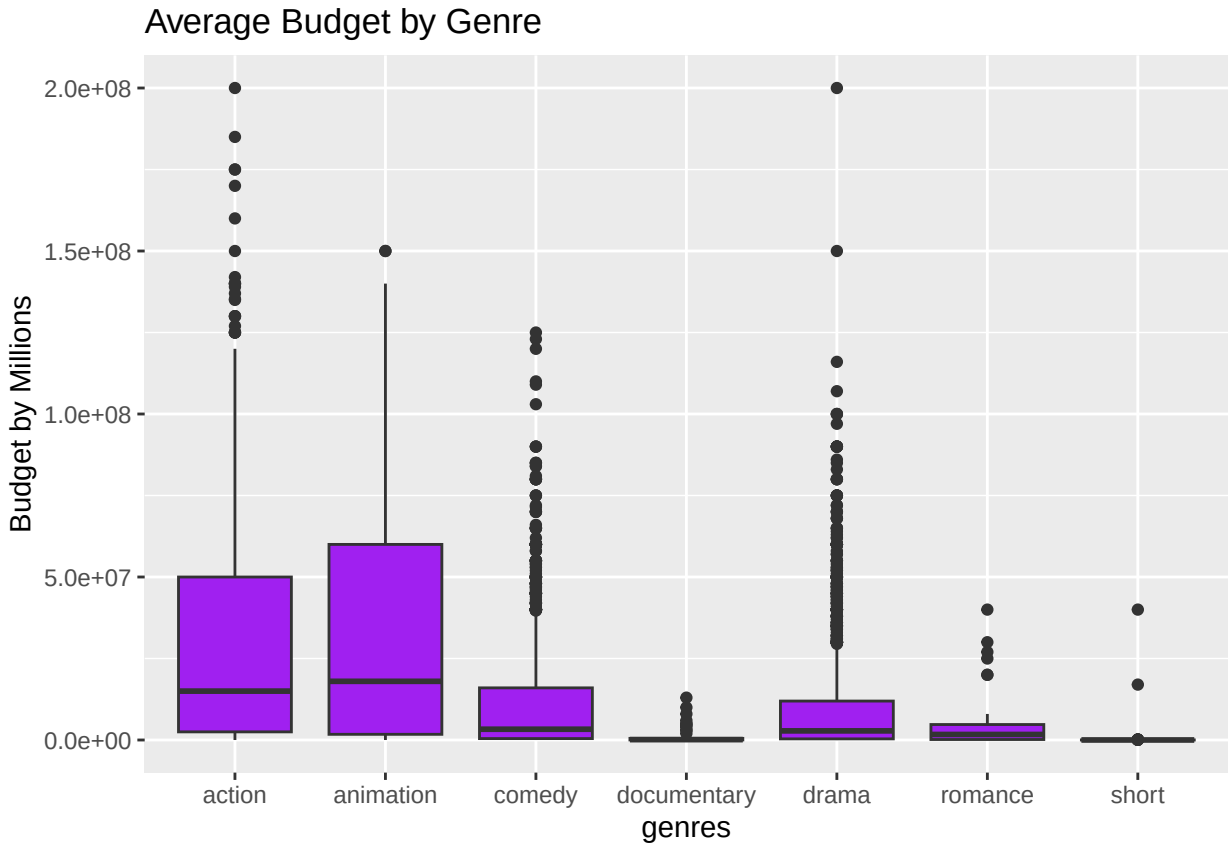
Question 9

Average Budget by Genres

Question : Which genres have the biggest budget comparing all 7 genres ?

Calculations are for budget >= 1000000

```
dfBudget <- dfGenres %>%
  group_by(genres) %>%
  filter(budget != "NA", budget <= 1000000) %>%
  mutate(budget_in_mil = budget/1000000) %>%
  summarize(avg_budget = mean(budget)) %>%
  invisible()

budgetPerM = dfBudget %>% pull(avg_budget)

  ggplot(data = dfGenres, aes(x=genres,y=budget))+
  geom_boxplot(fill="purple") +
  labs(y='Budget by Millions',title='Average Budget by Genre')
```

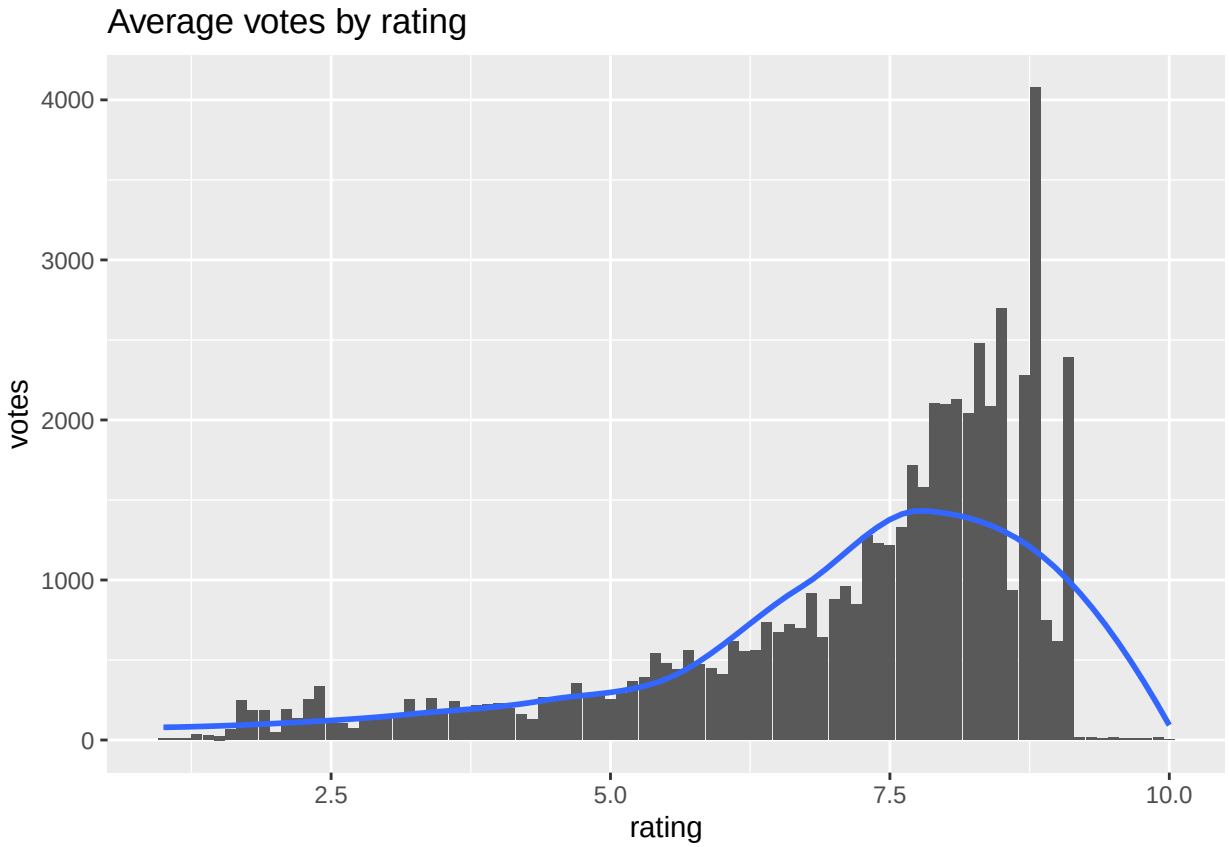


The graphs shows that action and animation genres have greater spending than all other genres.

Average votes by rating

Question : Is there a connection between number of votes and the rating of a movie?

```
dfvotes <- dfGenres %>%
  filter(!is.na(votes)) %>%
  group_by(rating) %>%
  summarize(mean_votes = mean(votes))

ggplot(data = dfvotes, aes(x = rating, y = mean_votes)) +
  geom_bar(stat = 'identity') +
  geom_smooth(method = "loess", se = F) +
  labs(x = 'rating', y='votes',title='Average votes by rating')
```



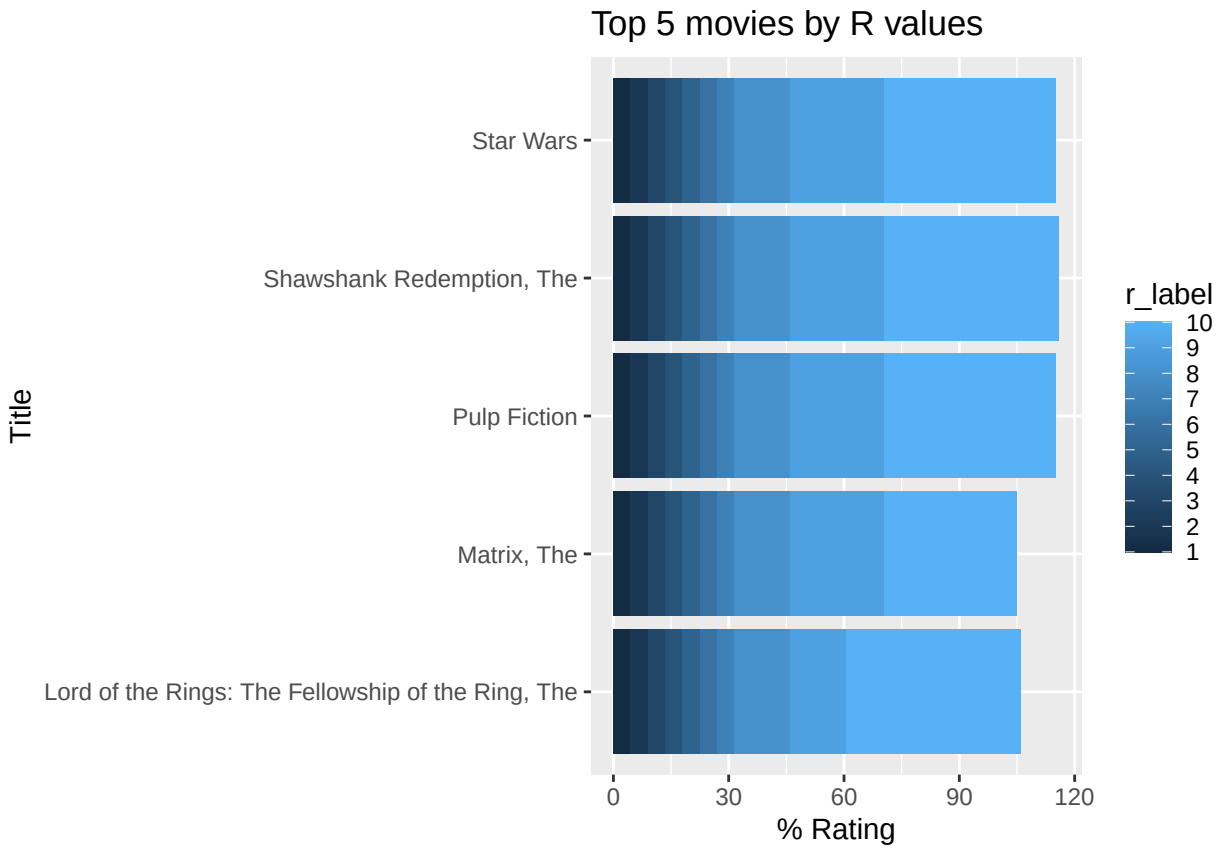
Looking at the distribution of votes among the rating we can observe the most voted movies are rated between 8.0 and 9.0. The number of movies that received more than 9.0 rating are scarce.

Top 5 voted movies by their rating values

Question : What percent of the ratings are given for the most voted movies ?

```
dfVoters <- dfGenres %>%
  arrange(desc(votes)) %>%
  top_n(5,votes)

dfr_main <- dfVoters
dfr_temp <- dfr_main %>%
  mutate(r_value = r1) %>%
  mutate(r_label = 1)
dfr_main_1 <- bind_rows(dfr_main, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r2) %>%
  mutate(r_label = 2)
dfr_main_2 <- bind_rows(dfr_main_1, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r3) %>%
  mutate(r_label = 3)
dfr_main_3 <- bind_rows(dfr_main_2, dfr_temp)

dfr_temp <- dfr_main %>%
```

```
  mutate(r_value = r4) %>%
  mutate(r_label = 4)
dfr_main_4 <- bind_rows(dfr_main_3, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r5) %>%
  mutate(r_label = 5)
dfr_main_5 <- bind_rows(dfr_main_4, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r6) %>%
  mutate(r_label = 6)
dfr_main_6 <- bind_rows(dfr_main_5, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r7) %>%
  mutate(r_label = 7)
dfr_main_7 <- bind_rows(dfr_main_6, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r8) %>%
  mutate(r_label = 8)
dfr_main_8 <- bind_rows(dfr_main_7, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r9) %>%
  mutate(r_label = 9)
dfr_main_9 <- bind_rows(dfr_main_8, dfr_temp)

dfr_temp <- dfr_main %>%
  mutate(r_value = r10) %>%
  mutate(r_label = 10)
dfr_main_10 <- bind_rows(dfr_main_9, dfr_temp)

dfr_main_10 <- dfr_main_10 %>%
  filter(!is.na(r_value))
##dfr_main_10

ggplot(data = dfr_main_10, aes(x=title, y = r_value, fill = r_label)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(y='% Rating',title='Top 5 movies by R values', x='Title')+
  scale_fill_gradient(breaks=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10))+
  coord_flip()
```



This plot shows that most people voted the same for the top 5 movies considering the first 8 percentile ratings. Interestingly except Lord of The Rings people voted more nines over tens.Lord of The Rings and Shawshank Redemption received most amount of tens among the top five voted movies.