Lab2

Tomohiro Justin Wada

2023-02-28

Question 1

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

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

```
min(movies$year)
```

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

Question 2

```
prop <- ((58788-sum(is.na(movies$budget)))/58788)*100
prop
```

```
## [1] 8.870858
```

```
naProp <- (sum(is.na(movies$budget))/58788)*100
naProp
```

```
## [1] 91.12914
```

```
topFive <- movies[order(movies$budget, decreasing = TRUE), ][1:5, ]
topFive$title
```

```
## [1] "Spider-Man 2"                          "Titanic"
## [3] "Troy"                                  "Terminator 3: Rise of the Machines"
## [5] "Waterworld"
```

8.87% of movies in the database have their budget included and 91.12% have no budget included in the database. The top 5 most expensive movie is “Spider-Man 2”, “Titanic”, “Troy”, “Terminator 3: Rise of the Machines”, and “Waterworld”.

Question 3

```
top5_longest <- movies[order(movies$length, decreasing = TRUE), ][1:5, ]
top5_longest$title
```

```
## [1] "Cure for Insomnia, The"
## [2] "Longest Most Meaningless Movie in the World, The"
## [3] "Four Stars"
## [4] "Resan"
## [5] "Out 1"
```

The top 5 longest movie is “Cure for Insomia”, “The Longest Most Meaningless Movies in the World”, “The Four Stars”, “Resan”, and “Out 1”

Question 4

```
short <- movies[movies$Short == 1,]
s <- short[which.min(short$length),]
s
```

```
## # A tibble: 1 x 24
##   title      year length budget rating votes    r1    r2    r3    r4    r5    r6
##   <chr>     <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 17 Secon~  1998      1     NA    5.1     7     0     0     0  14.5  24.5  14.5
## # ... with 12 more variables: r7 <dbl>, r8 <dbl>, r9 <dbl>, r10 <dbl>,
## #   mpaa <chr>, Action <int>, Animation <int>, Comedy <int>, Drama <int>,
## #   Documentary <int>, Romance <int>, Short <int>
```

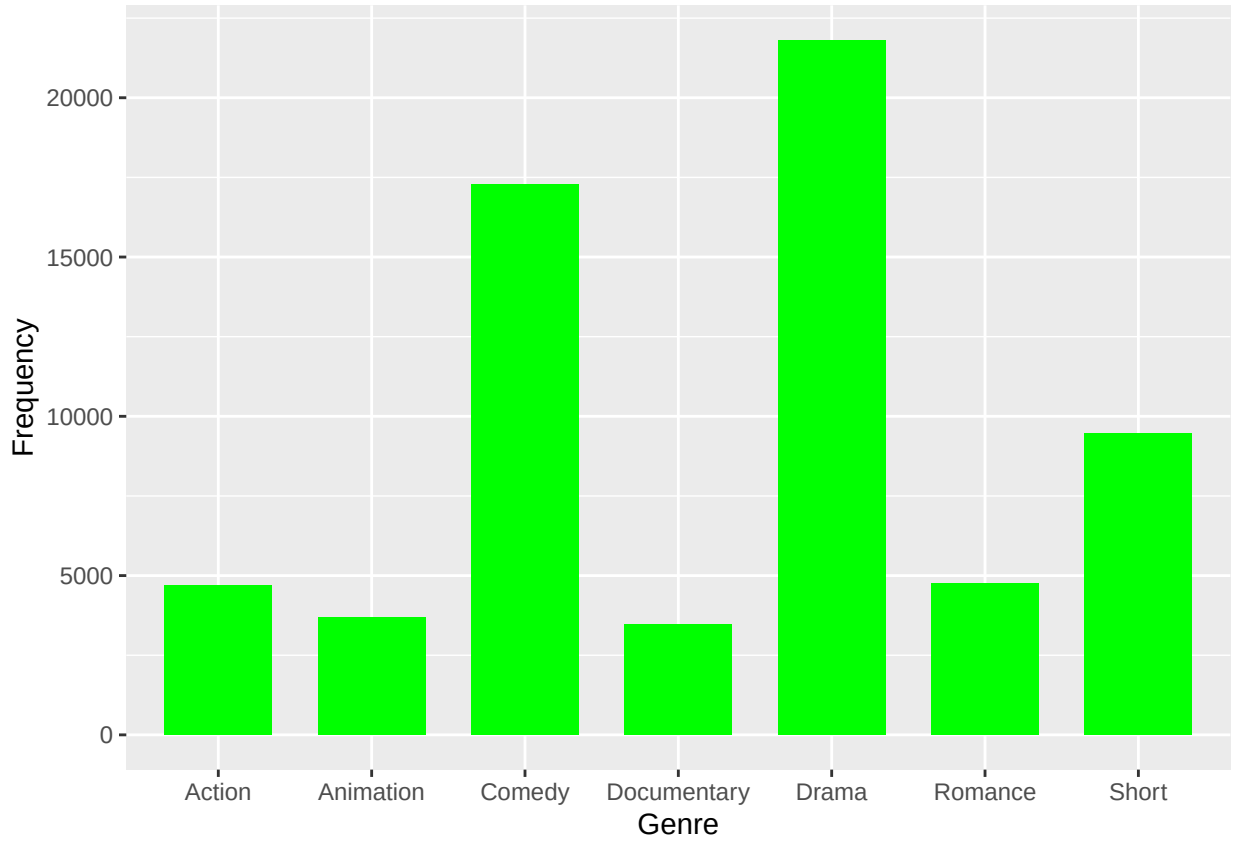
```
short <- movies[movies$Short == 1,]
s <- short[which.max(short$length),]
s
```

```
## # A tibble: 1 x 24
##   title      year length budget rating votes    r1    r2    r3    r4    r5    r6
##   <chr>     <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 10 jaar ~  2004    240     NA    7.9     7     0     0  14.5     0     0  14.5
## # ... with 12 more variables: r7 <dbl>, r8 <dbl>, r9 <dbl>, r10 <dbl>,
## #   mpaa <chr>, Action <int>, Animation <int>, Comedy <int>, Drama <int>,
## #   Documentary <int>, Romance <int>, Short <int>
```

The shortest length movie is 17 Seconds to Sophie and the longest is 10 jaar leuven kort

Question 5

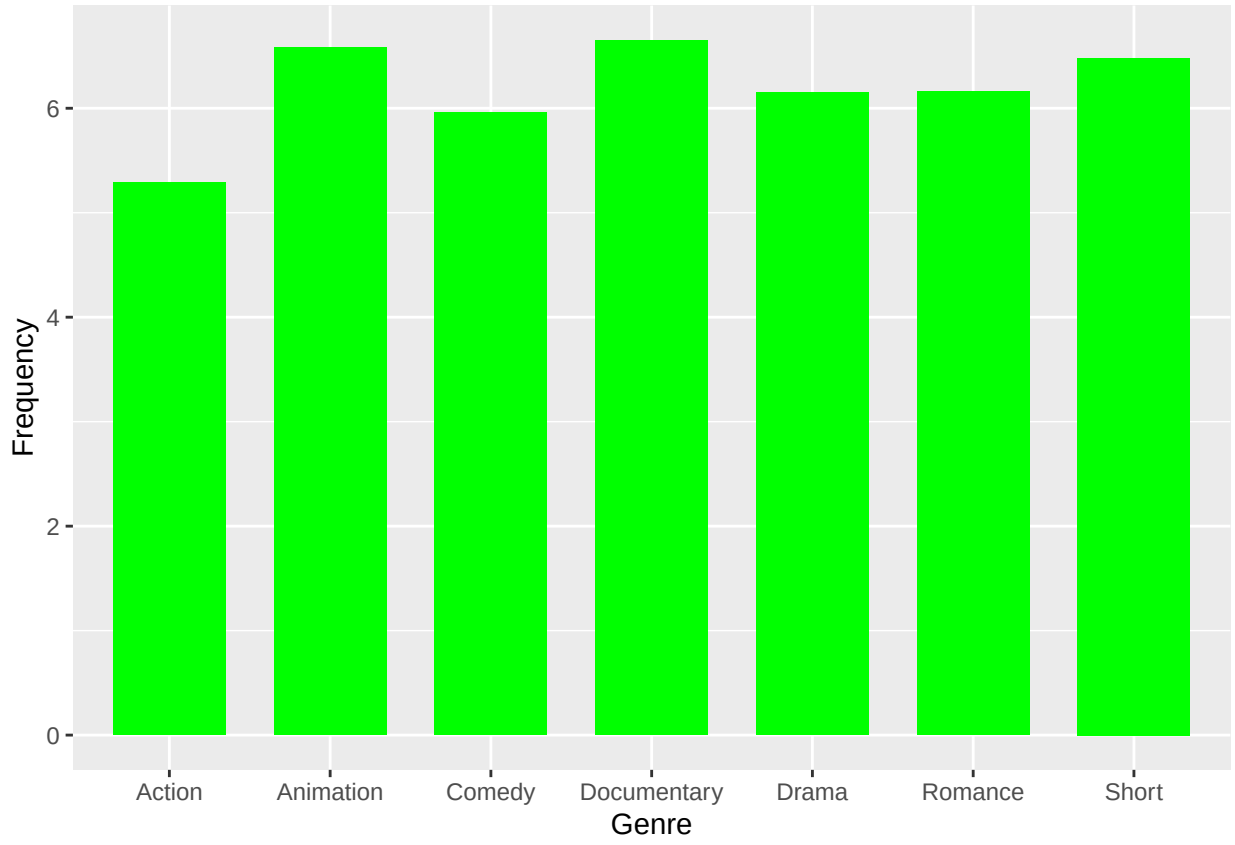
```
movGenre <- data.frame(Genre = c("Action", "Animation", "Comedy", "Drama", "Documentary","Romance","Sho

ggplot(movGenre, aes(x = Genre, y = Frequency))+ geom_bar(stat = "Identity", width= 0.7, fill="green")
```



Question 6

```
action <- movies[movies$Action == 1,]
avgAction <- mean(action$rating)

animation <- movies[movies$Animation == 1,]
avgAnimation <- mean(animation$rating)

comedy <- movies[movies$Comedy == 1,]
avgComedy <- mean(comedy$rating)

drama <- movies[movies$Drama == 1,]
avgDrama <- mean(drama$rating)

documentary <- movies[movies$Documentary == 1,]
avgDocumentary <- mean(documentary$rating)

romance <- movies[movies$Romance == 1,]
```

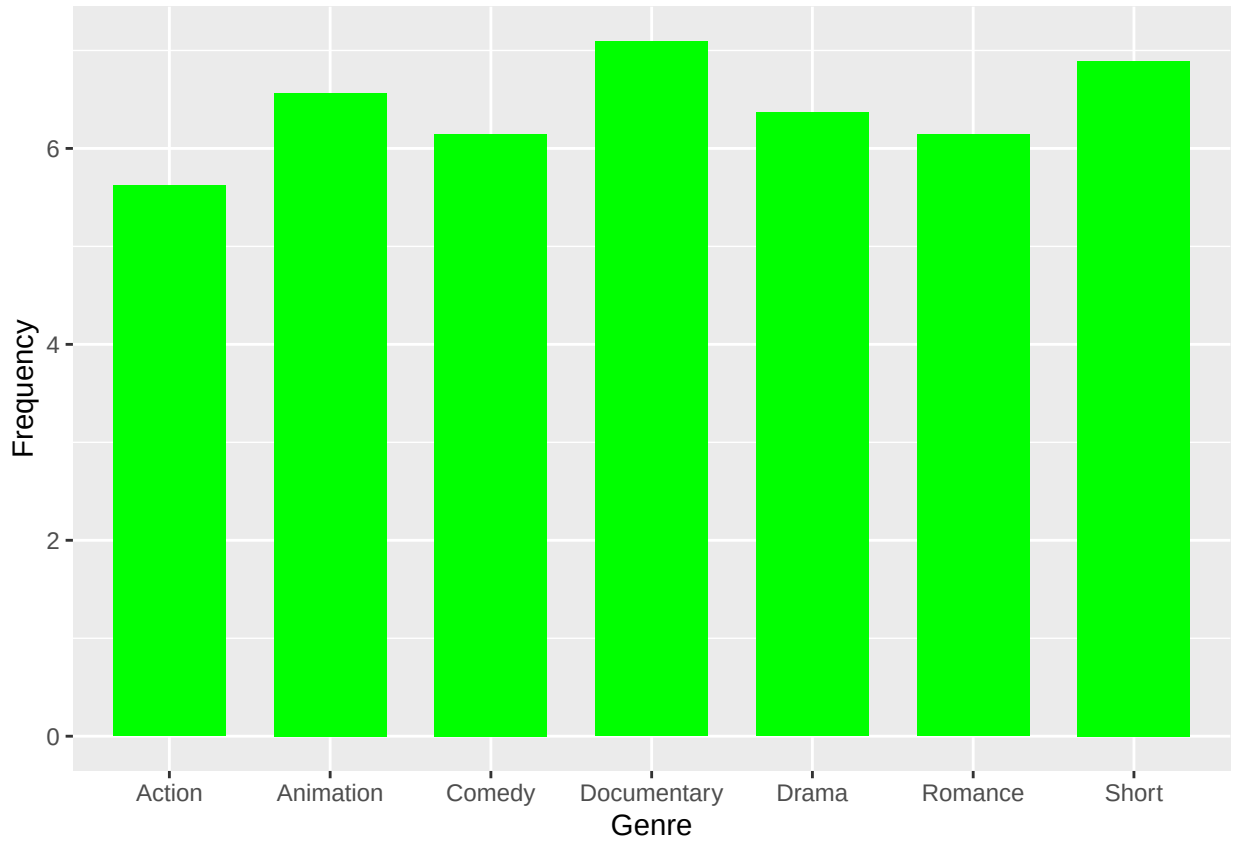
```
avgRomance <- mean(romance$rating)

short <- movies[movies$Short == 1,]
avgShort <- mean(short$rating)

movGenre <- data.frame(Genre = c("Action", "Animation", "Comedy", "Drama", "Documentary","Romance","Sho

ggplot(movGenre, aes(x = Genre, y = Frequency))+ geom_bar(stat = "Identity", width= 0.7, fill="green")
```



Question 7

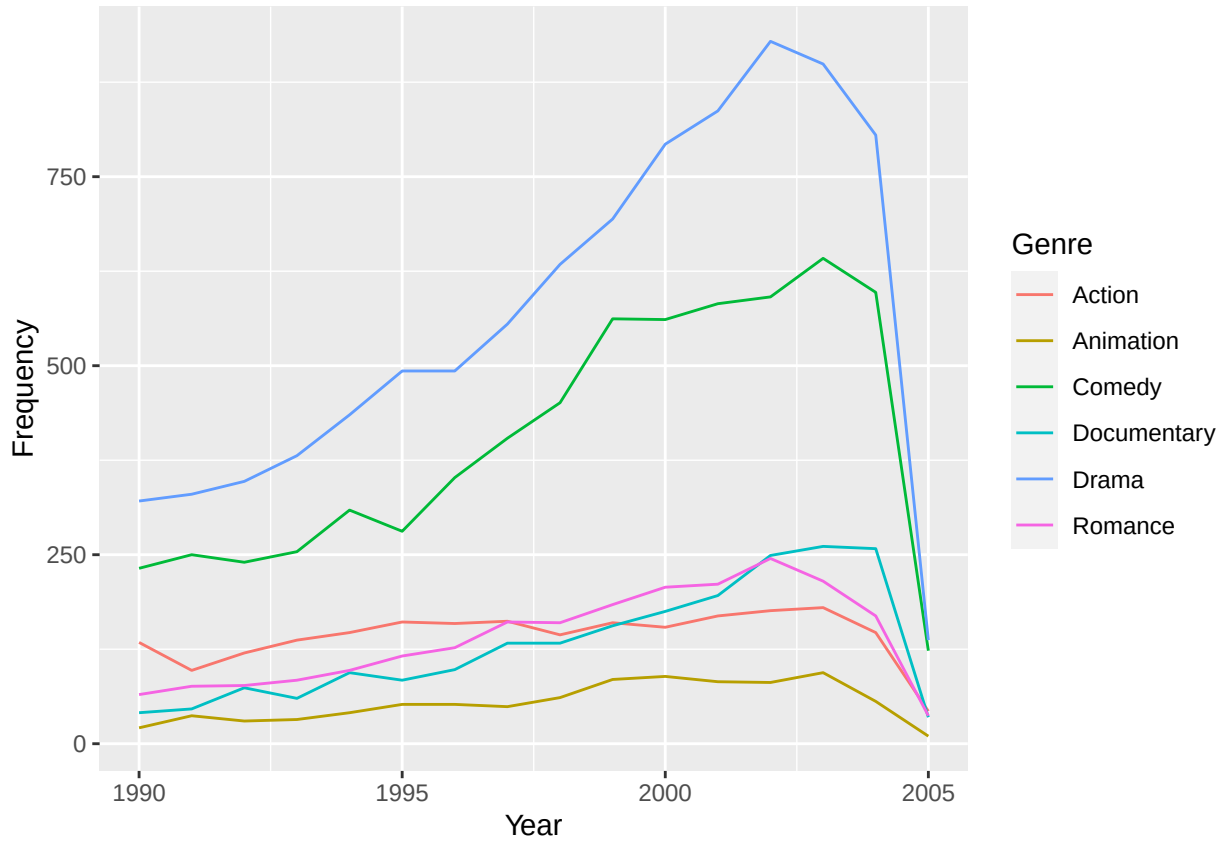
```
action <- movies[movies$Action == 1,]
actionYear <- action[action$year >= 2000 & action$year <= 2005,]
avgAction <- mean(actionYear$rating)

animation <- movies[movies$Animation == 1,]
animationYear <- animation[animation$year >= 2000 & animation$year <= 2005,]
avgAnimation <- mean(animationYear$rating)

comedy <- movies[movies$Comedy == 1,]
comedyYear <- comedy[comedy$year >= 2000 & comedy$year <= 2005,]
avgComedy <- mean(comedyYear$rating)
```

```r
drama <- movies[movies$Drama == 1,]
dramaYear <- drama[drama$year >= 2000 & drama$year <= 2005,]
avgDrama <- mean(dramaYear$rating)

documentary <- movies[movies$Documentary == 1,]
documentaryYear <- documentary[documentary$year >= 2000 & documentary$year <= 2005,]
avgDocumentary <- mean(documentaryYear$rating)

romance <- movies[movies$Romance == 1,]
romanceYear <- romance[romance$year >= 2000 & romance$year <= 2005,]
avgRomance <- mean(romanceYear$rating)

short <- movies[movies$Short == 1,]
shortYear <- short[short$year >= 2000 & short$year <= 2005,]
avgShort <- mean(shortYear$rating)

movGenre <- data.frame(Genre = c("Action", "Animation", "Comedy", "Drama", "Documentary","Romance","Sho

ggplot(movGenre, aes(x = Genre, y = Frequency))+ geom_bar(stat = "Identity", width= 0.7, fill="green")
```



Question 8

```
Year <- 1990:2005
movie <- movies[movies$year >= 1990,]

ActionMovieYear <- movie[movie$Action == 1,]
AnimationMovieYear <- movie[movie$Animation == 1,]
ComedyMovieYear <- movie[movie$Comedy == 1,]
DramaMovieYear <- movie[movie$Drama == 1,]
DocumentaryMovieYear <- movie[movie$Documentary == 1,]
RomanceMovieYear <- movie[movie$Romance == 1,]

MovieGenre <- data.frame(Year = Year)

MovieGenre <- MovieGenre %>%
  group_by(Year) %>%
  mutate(Action = sum(ActionMovieYear$year == Year),
         Animation = sum(AnimationMovieYear$year == Year),
         Comedy = sum(ComedyMovieYear$year == Year),
         Drama = sum(DramaMovieYear$year == Year),
         Documentary = sum(DocumentaryMovieYear$year == Year),
         Romance = sum(RomanceMovieYear$year == Year)) %>%
  gather(key = Genre, value = Frequency, -Year)


ggplot(data = MovieGenre, aes(x = Year, y = Frequency, color = Genre)) +
  geom_line() +
  labs(x = "Year", y = "Frequency", color = "Genre")
```



Question 9

1) Which movie has the highest budget?

```
max(movies$budget, na.rm=TRUE)
```

```
## [1] 200000000
```

The highest movie budget is 200000000.

2) What is the number of movies that have a budget greater than $1 million and have a rating greater than 5?

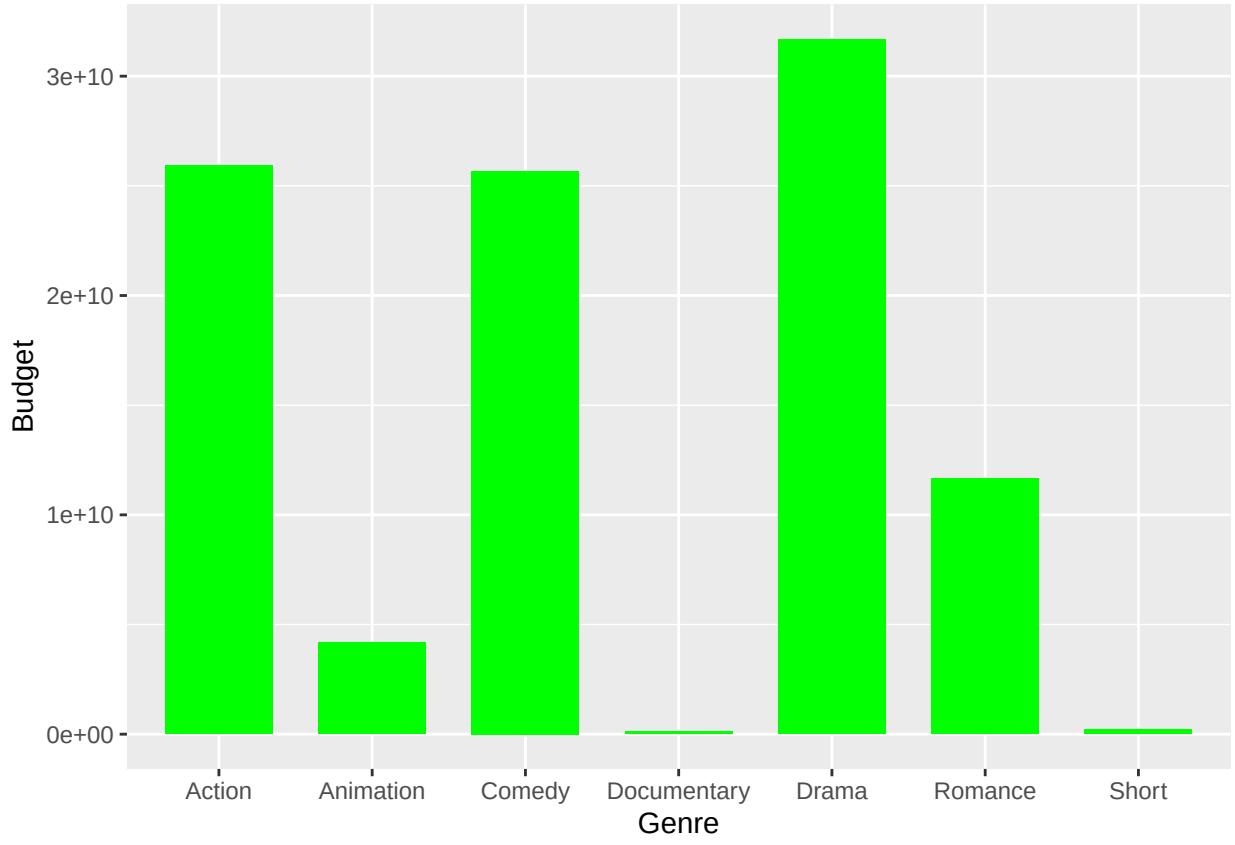
```
sum(movies$budget > 1000000 & movies$rating > 5, na.rm=T)
```

```
## [1] 2506
```

The number of movies that have a budget greater than $1 million and have a rating greater than 5 is 2506 movies.

3) Plot each movie genre budget. Which genre has the highest budget?

```r
action <- movies[movies$Action == 1,]
avgAction <- sum(action$budget, na.rm=T)

animation <- movies[movies$Animation == 1,]
avgAnimation <- sum(animation$budget, na.rm=T)

comedy <- movies[movies$Comedy == 1,]
avgComedy <- sum(comedy$budget, na.rm=T)

drama <- movies[movies$Drama == 1,]
avgDrama <- sum(drama$budget, na.rm=T)

documentary <- movies[movies$Documentary == 1,]
avgDocumentary <- sum(documentary$budget, na.rm=T)

romance <- movies[movies$Romance == 1,]
avgRomance <- sum(romance$budget, na.rm=T)

short <- movies[movies$Short == 1,]
avgShort <- sum(short$budget, na.rm=T)

movGenre <- data.frame(Genre = c("Action", "Animation", "Comedy", "Drama", "Documentary","Romance","Sho

ggplot(movGenre, aes(x = Genre, y = Budget))+ geom_bar(stat = "Identity", width= 0.7, fill="green")
```



Drama has the highest the movie budget.