Lab2 Saunders

Taylor Saunders

2023-11-05

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(ggplot2movies)
```

```
## Warning: package 'ggplot2movies' was built under R version 4.3.1
```

```
data(movies)
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## 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
```

Question 1

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

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

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

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

```
print("The range of years of production of the movies of this dataset is 1893 to 2005" )
```

```
## [1] "The range of years of production of the movies of this dataset is 1893 to 2005"
```

Question 2

```
#part 1
has_budget <- sum(is.na(movies$budget))
print(has_budget)
```

```
## [1] 53573
```

```
#dim(movies) #checking values are reasonable
no_budget <- sum(!is.na(movies$budget))
print(no_budget)
```

```
## [1] 5215
```

```
has_budget_percent <- 53573 / 58788  * 100
print(has_budget_percent)
```

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

```
no_budget_percent <- 5215 / 58788 * 100
print(no_budget_percent)
```

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

91% of the movies had a value for their budget, while 9% of the movies did not have a listed budget

```
#part 2
top_5_expense <- arrange(movies, desc(budget)) #arrange in descending order based on length value
head(top_5_expense, n = 5)
```

```
## # A tibble: 5 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 Spider-M~  2004    127 2   e8    7.9 40256   4.5   4.5   4.5   4.5   4.5   4.5
## 2 Titanic    1997    194 2   e8    6.9 90195  14.5   4.5   4.5   4.5   4.5   4.5
## 3 Troy       2004    162 1.85e8    7.1 33979   4.5   4.5   4.5   4.5   4.5  14.5
## 4 Terminat~  2003    109 1.75e8    6.9 32111   4.5   4.5   4.5   4.5   4.5  14.5
## 5 Waterwor~  1995    176 1.75e8    5.4 19325   4.5   4.5   4.5  14.5  14.5  14.5
## # i 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 top 5 most expensive movies in this data set are Spider-Man 2 and the Titanic, both with the same budget of 200000000, Troy, Terminator 3: Rise of the Machines, and Waterworld.

Question 3

```
top_5 <- arrange(movies, desc(length)) #arrange in descending order based on length value to see highes
head(top_5, n = 5)
```

```
## # A tibble: 5 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 Cure for~  1987   5220     NA    3.8    59  44.5   4.5   4.5   4.5   0     0
## 2 Longest ~  1970   2880     NA    6.4    15  44.5   0     0     0     0     0
## 3 Four Sta~  1967   1100     NA    3      12  24.5   0     4.5   0     0     0
## 4 Resan      1987    873     NA    5.5    12   0     0     4.5   0     0     0
## 5 Out 1      1971    773     NA    6.7    20   4.5   4.5   4.5   0     4.5  14.5
## # i 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 top 5 longest movies are The Cure for Insomnia, The Longest Most Meaningless Movie in the World, Four Stars, Resan, and Out 1.

Question 4

```
shortest_movie <- arrange(movies, length) #sort in ascending order
head(shortest_movie)
```

```
## # A tibble: 6 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
## 2 2 A.M. i~  1905      1     NA    5.2    13   0     0    14.5  14.5  24.5  34.5
## 3 Admiral ~  1897      1     NA    4.4    34   4.5   4.5   4.5  14.5  14.5  14.5
## 4 Admiral ~  1899      1     NA    4.1    27  14.5   4.5  24.5   4.5  24.5  14.5
## 5 Alphonse~  1903      1     NA    4.1     9   0     0    34.5  14.5  44.5  14.5
## 6 Ameta      1903      1     NA    4.9    11   0     4.5  14.5   4.5  14.5  44.5
## # i 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>
```

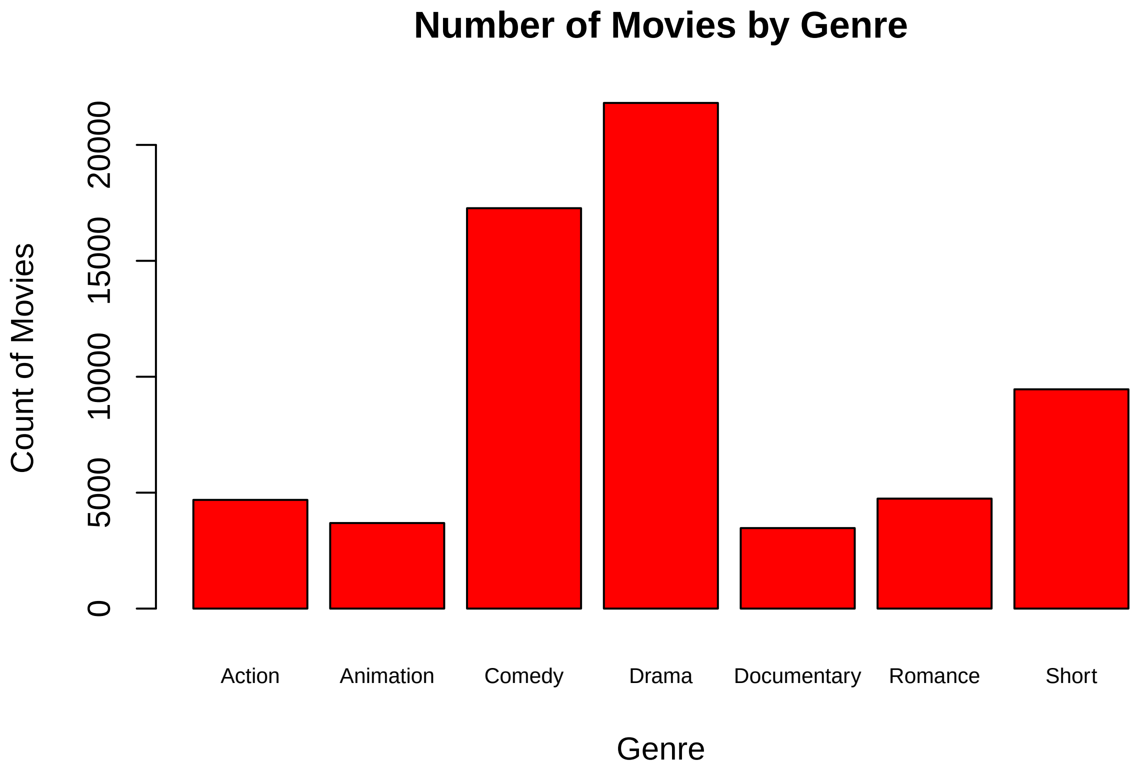
```
table(shortest_movie$length == 1) #more than one movie has the shortest length
```

```
##
## FALSE  TRUE
## 58619   169
```

The head of the first 5 shortest movies is 17 seconds to Sophie, 2 A.M. in the Subway, Admiral Cigarette, Admiral Dewey Leading Land Parade, Alphonse and Gaston No. 3. However, there are 169 movies in this data set with the length "1".

Question 5

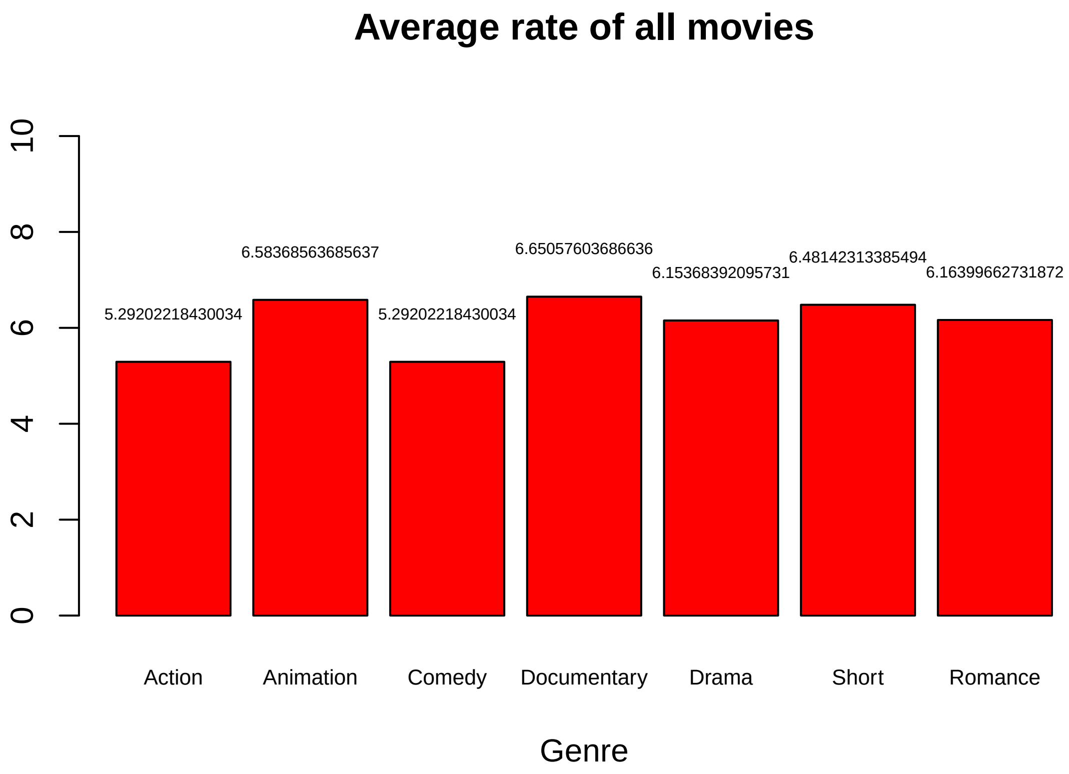
```
genre <- movies %>% #dataset being used
  select(Action, Animation, Comedy, Drama, Documentary, Romance, Short) %>% #the genres for our plot
  colSums() #sum of each column

barplot(genre,
        main = "Number of Movies by Genre",
        xlab = "Genre",
        ylab = "Count of Movies",
        cex.names=0.7,
        col = "red"
)
```



Question 6

```
action = filter(movies, Action == 1)
actionrating = mean(action$rating)

animation = filter(movies, Animation == 1)
animationrating = mean(animation$rating)

comedy = filter(movies, Comedy == 1)
comedyrating = mean(action$rating)

drama = filter(movies, Drama == 1)
dramarating = mean(drama$rating)

Doc = filter(movies, Documentary == 1)
docrating = mean(Doc$rating)

shorts = filter(movies, Short == 1)
shortrating = mean(shorts$rating)

romance = filter(movies, Romance == 1)
romancerating = mean(romance$rating)

averagerate <- c(actionrating, animationrating, comedyrating, docrating,
              dramarating, shortrating, romancerating)

counts = as.vector(averagerate)
```

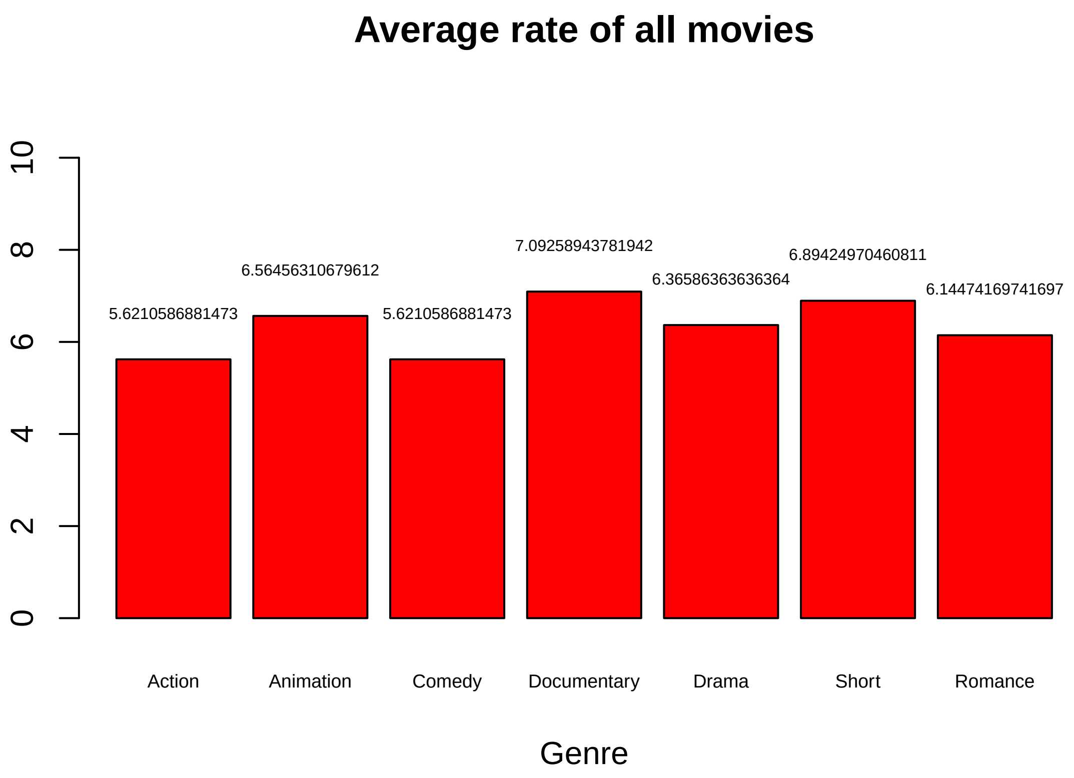
```
xx <-
  barplot(
    averagerate,
    main = "Average rate of all movies",
    names = c('Action','Animation','Comedy','Documentary','Drama','Short','Romance'),
    xlab = "Genre",
    ylim = c(0, max(averagerate) + 4),
    cex.names=0.7,
    col = "Red"
  )
text(
  x <- xx,
  y = averagerate + 1,
  label = as.character(averagerate),
  cex = 0.5,
  col = "Black"
)
```



Question 7

```
action = filter(movies, Action == 1, year >= 2000 & year <= 2005)
actionrating = mean(action$rating)

animation = filter(movies, Animation == 1, year >= 2000 & year <= 2005)
animationrating = mean(animation$rating)

comedy = filter(movies, Comedy == 1, year >= 2000 & year <= 2005)
```

```
comedyrating = mean(action$rating)

drama = filter(movies, Drama == 1, year >= 2000 & year <= 2005)
dramarating = mean(drama$rating)

Doc = filter(movies, Documentary == 1, year >= 2000 & year <= 2005)
docrating = mean(Doc$rating)

shorts = filter(movies, Short == 1, year >= 2000 & year <= 2005)
shortrating = mean(shorts$rating)

romance = filter(movies, Romance == 1, year >= 2000 & year <= 2005)
romancerating = mean(romance$rating)

averagerate <- c(actionrating, animationrating, comedyrating, docrating,
              dramarating, shortrating, romancerating)

counts = as.vector(averagerate)
xx <-
  barplot(
    averagerate,
    main = "Average rate of all movies",
    names = c('Action','Animation','Comedy','Documentary','Drama','Short','Romance'),
    xlab = "Genre",
    ylim = c(0, max(averagerate) + 4),
    cex.names=0.6,
    col = "Red"
  )
text(
  x <- xx,
  y = averagerate + 1,
  label = as.character(averagerate),
  cex = 0.5,
  col = "Black"
)
```



Question 8

```
selected_genres <- movies %>%
  select(Action, Animation, Comedy, Drama, Documentary, Romance, year) %>%
  filter(year >= 1990)

plot_top6 <- function(genre){
  plot_genres <- selected_genres %>%
  group_by(year) %>%
  summarise(
    action = sum(Action),
    animation = sum(Animation),
    comedy = sum(Comedy),
    drama = sum(Drama),
    documentary= sum(Documentary),
    romance = sum(Romance),
  )

}
print(plot_top6())
```

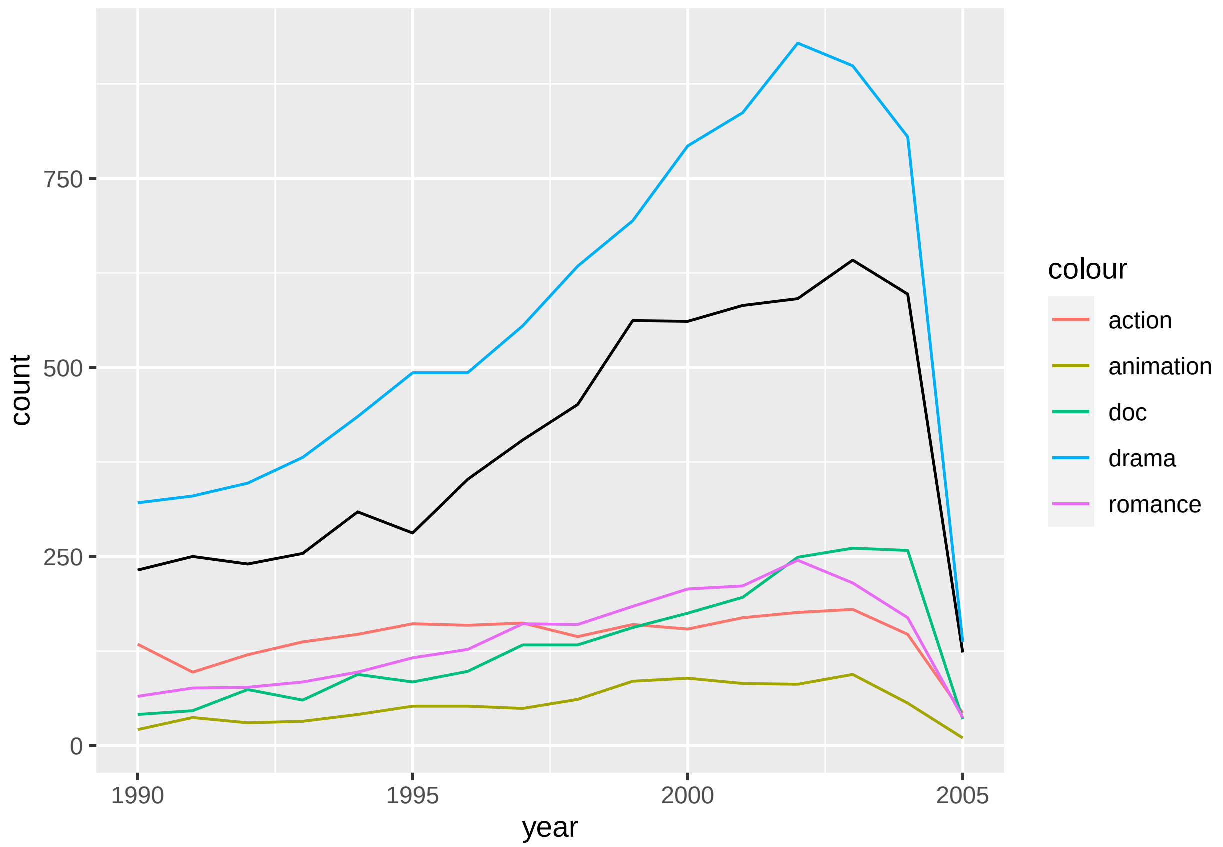
```
## # A tibble: 16 x 7
##     year action animation comedy drama documentary romance
##    <int>  <int>     <int>  <int> <int>       <int>   <int>
##  1  1990    134        21    232   321          41      65
##  2  1991     97        37    250   330          46      76
##  3  1992    120        30    240   347          74      77
##  4  1993    137        32    254   381          60      84
##  5  1994    147        41    309   435          94      97
##  6  1995    161        52    281   493          84     116
##  7  1996    159        52    352   493          98     127
##  8  1997    162        49    404   555         133     161
##  9  1998    144        61    451   634         133     160
## 10  1999    160        85    562   694         156     184
## 11  2000    154        89    561   793         175     207
## 12  2001    169        82    582   837         196     211
## 13  2002    176        81    591   929         249     245
## 14  2003    180        94    642   899         261     215
## 15  2004    147        56    597   805         258     169
## 16  2005     43        10    123   137          35      37
```

Comment: When using a function, the appropriate tables are made. However, I am uncertain as to how to graph them using the function.

Plotting without the use of a function:

```
plot_genre <-
  ggplot2movies::movies %>%
  filter(year >= 1990) %>%
  select(Action, Animation, Comedy, Drama, Documentary, Romance, year) %>%
  group_by(year) %>%
  summarise(
    action = sum(Action),
    animation = sum(Animation),
    comedy = sum(Comedy),
    drama = sum(Drama),
    documentary = sum(Documentary),
    romance = sum(Romance),
  )
 ggplot(plot_genre, aes(x = year)) +
  geom_line(aes(y = action, color = "action")) +
  geom_line(aes(y = animation, color = "animation")) +
  geom_line(aes(y = comedy, colur = "comedy")) +
  geom_line(aes(y = drama, color = "drama")) +
  geom_line(aes(y = documentary, color = "doc")) +
  geom_line(aes(y = romance, color = "romance")) +
  ylab("count")
```

```
## Warning in geom_line(aes(y = comedy, colur = "comedy")): Ignoring unknown
## aesthetics: colur
```



Question 9 1. How many movies were published in 2002? 2. What is the top movie for the Action Genre in 2002? 3. What is the top movie for the Action Genre in 2002?

```
#1
#3
movies_2002 <- sum(movies$year == "2002")
print(movies_2002)
```

```
## [1] 2168
```

There were 2,168 movies published in 2002

```
#2
top_action <- filter(movies, Action == 1, year == 2002)
top_action <- arrange(top_action, desc(rating))
head(top_action)
```

```
## # A tibble: 6 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 Sundown  2002     19 NA         9.5      6   0     0     0     0     0     0
## 2 More T~  2002     70 NA         9.4      5   0     0     0     0     0     0
## 3 Suspen~  2002     21 NA         9.3     12   0     0     0     0     4.5   0
## 4 Enrage~  2002     50  3.5 e3    8.9     25  24.5   4.5   0     0     0     4.5
## 5 Lord o~  2002    223  9.40e7    8.8 114797   4.5   4.5   4.5   4.5   4.5   4.5
## 6 Outdoo~  2002     64 NA         8.8     15   0     0     0     0     0     0
## # i 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 top rated action movie in 2002 was Sundown, with a rating of 9.5.

```
#3
top_drama <- filter(movies, Drama == 1, year == 2002)
head(top_drama)
```

```
## # A tibble: 6 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 $windle    2002     93     NA    5.3   200   4.5   0     4.5   4.5  24.5  24.5
## 2 (A)Torzi~  2002     13     NA    7.2    71   4.5   0     4.5   4.5   4.5   4.5
## 3 (Entre n~  2002     82     NA    4.8    22  14.5   4.5   4.5  14.5  24.5  14.5
## 4 *Corpus ~  2002     92     NA    4.9    36  24.5  14.5   4.5   4.5   0     4.5
## 5 11'09''0~  2002    134     NA    6.9  1264   4.5   4.5   4.5   4.5   4.5  14.5
## 6 12:35      2002     92     NA    8.2     5  24.5   0     0     0     0     0  
## # i 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>
```

```
top_drama <- arrange(top_drama, desc(rating))
head(top_drama)
```

```
## # A tibble: 6 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 Mutual A~  2002     89     NA    9.6     9     0   0         0     0     0     0
## 2 Sundown    2002     19     NA    9.5     6     0   0         0     0     0     0
## 3 More Tha~  2002     70     NA    9.4     5     0   0         0     0     0     0
## 4 Dusk       2002     33     NA    9.3    14     0   4.5       0     0     0     0
## 5 Half Sis~  2002      3     NA    9.3     6     0   0         0     0     0     0
## 6 Unborn     2002      8     NA    9.3     6     0   0         0     0     0     0
## # i 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>
```

```
dim(top_drama)
```

```
## [1] 929  24
```

The top rated drama movie in 2002 was Mutual Admiration Society, with a rating of 9.6.