

Lab 2 Assignment

Grina Hwang

2023-11-05

Question 1

What is the year of production of the oldest movie and of the most recent movie in this data set

```
range(movies$year)
```

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

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

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

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

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

Question 2

```
# Proportion of movies that have their budget included in this database
```

```
rm.na.budget <- sum(na.omit(movies$budget))
```

```
print(rm.na.budget)
```

```
## [1] 69946256597
```

```
# Proportion of movies that don't have their budget included in this database
```

```
na.budget <- sum(is.na(movies$budget))
```

```
print(na.budget)
```

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

```
# Top 5 most expensive movies in this data set
```

```
topfive <- movies %>% arrange(desc(budget)) %>% select(title, budget)
```

```
head(topfive, 5)
```

```
## # A tibble: 5 x 2
```

```
##   title                                budget
```

```
##   <chr>                                <int>
```

```
## 1 Spider-Man 2                        200000000
```

```
## 2 Titanic                             200000000
```

```
## 3 Troy                                185000000
```

```
## 4 Terminator 3: Rise of the Machines 175000000
```

```
## 5 Waterworld                          175000000
```

Question 3

```
# Top 5 longest movies
topfivelongest <- movies %>% arrange(desc(length)) %>% select(title, length)
head(topfivelongest, 5)
```

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

Question 4

```
# Shortest movie and its length in minutes
shortestmovie = filter(movies, Short == 1, length == 1) %>% select(title, length)
head(shortestmovie, 1)
```

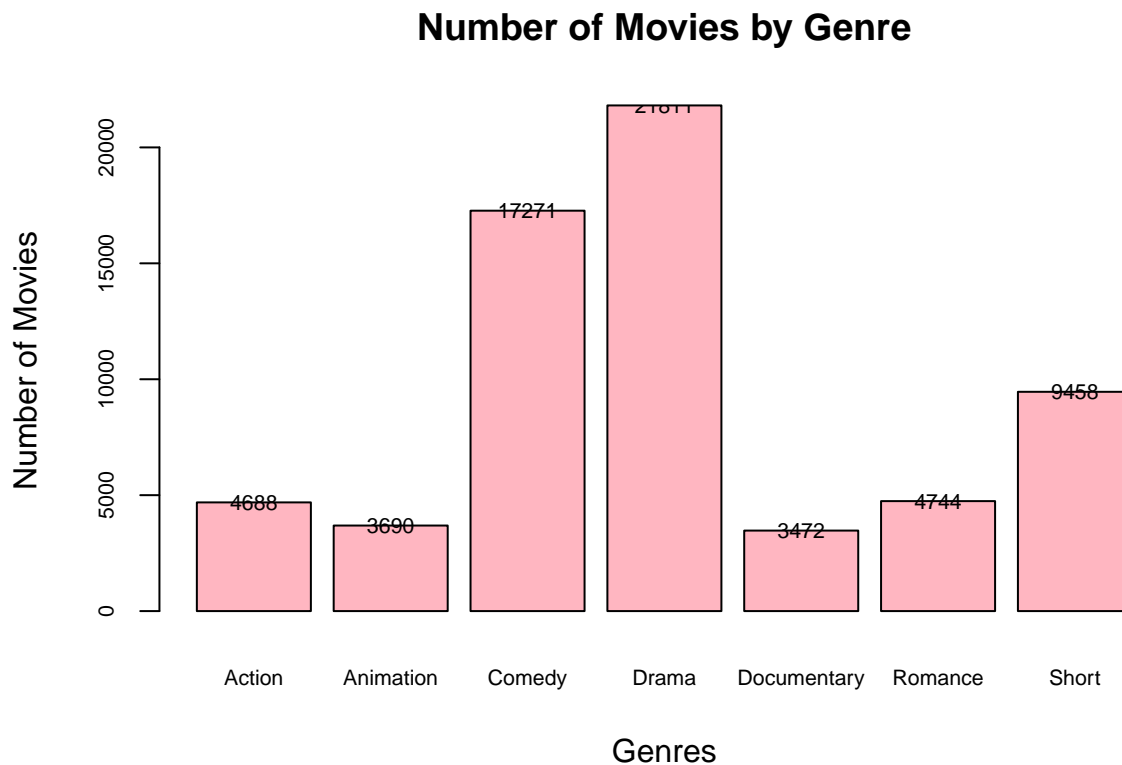
```
## # A tibble: 1 x 2
##   title                                length
##   <chr>                                <int>
## 1 17 Seconds to Sophie                1
```

```
# Longest movie and its length in minutes
longestmovie <- movies %>% arrange(desc(length)) %>% select(title, length)
head(longestmovie, 1)
```

```
## # A tibble: 1 x 2
##   title                                length
##   <chr>                                <int>
## 1 Cure for Insomnia, The              5220
```

Question 5

```
moviedatabase <- movies %>%
  select(Action, Animation, Comedy, Drama, Documentary, Romance, Short) %>%
  colSums()
xx <- barplot(moviedatabase,
  main = "Number of Movies by Genre",
  xlab = "Genres",
  ylab = "Number of Movies",
  cex.axis=0.7, cex.names=0.7,
  col = "light pink")
text(
  x = xx,
  y = moviedatabase + 1,
  label = as.character(moviedatabase),
  cex = 0.7,
  col = "Black"
)
```



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)

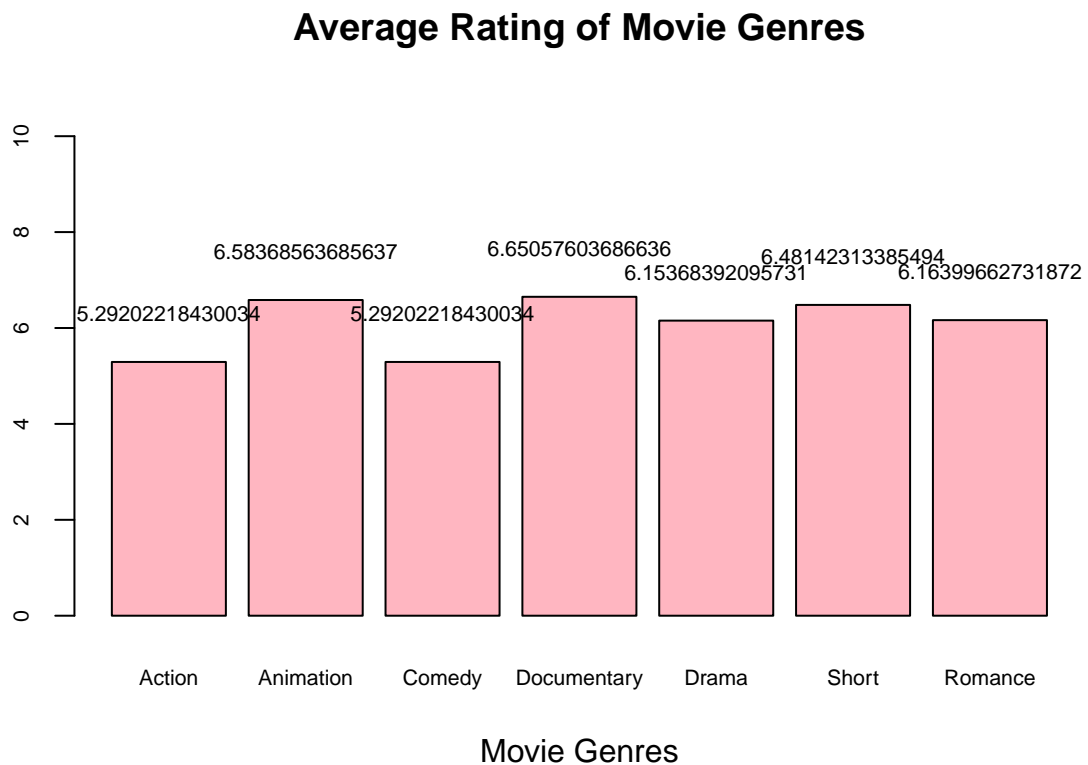
averagerate <- c(actionrating, animationrating, comedyrating, docrating,
```

```

dramarating, shortrating, romancerating)

counts = as.vector(averagerate)
xx <-
  barplot(
    averagerate,
    main = "Average Rating of Movie Genres",
    names = c('Action', 'Animation', 'Comedy', 'Documentary', 'Drama', 'Short', 'Romance'),
    xlab = "Movie Genres",
    ylim = c(0, max(averagerate) + 4),
    cex.axis=0.7, cex.names=0.7,
    col = "light pink"
  )
text(
  x = xx,
  y = averagerate + 1,
  label = as.character(averagerate),
  cex = 0.7,
  col = "Black"
)

```



Question 7

```

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)

filter(movies, year >= 2000 & year <= 2005)

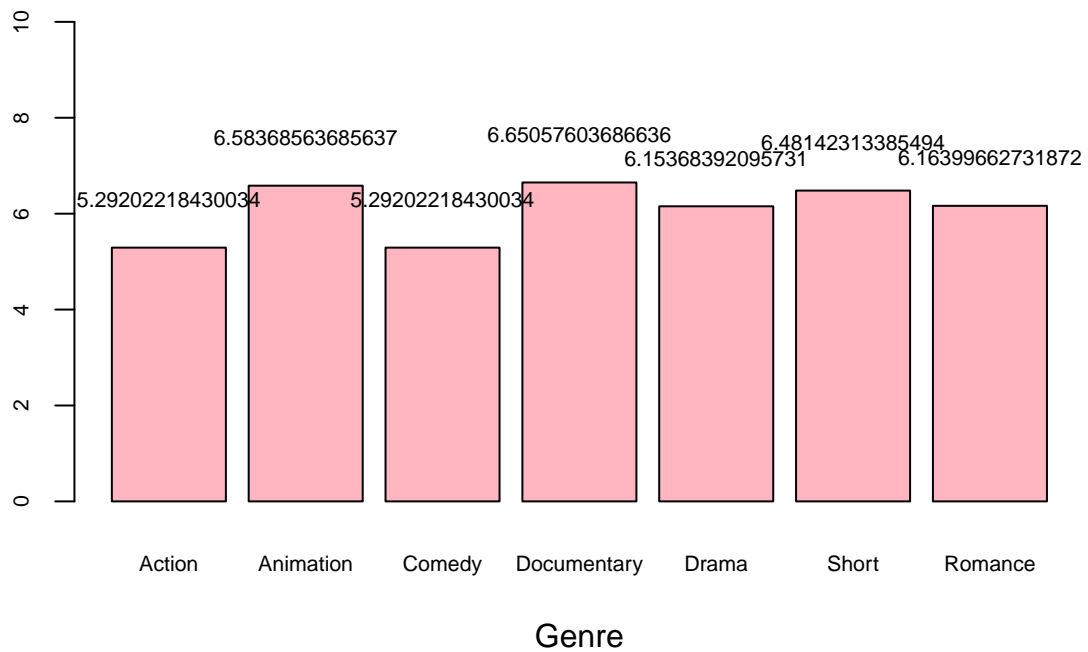
## # A tibble: 10,789 x 24
##   title    year length budget rating votes   r1    r2    r3    r4    r5    r6
##   <chr>   <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 $pent    2000     91    NA    4.3   45   4.5   4.5   4.5  14.5  14.5  14.5
## 2 $windle  2002     93    NA    5.3  200   4.5    0   4.5   4.5  24.5  24.5
## 3 '15'     2002     25    NA    6.7   24   4.5   4.5   4.5   4.5   4.5  14.5
## 4 'R Xmas  2001     83    NA    4.9  288  14.5   4.5   4.5   4.5  14.5  24.5
## 5 (A)Torz~ 2002     13    NA    7.2   71   4.5    0   4.5   4.5   4.5   4.5
## 6 (Entre ~ 2002     82    NA    4.8   22  14.5   4.5   4.5  14.5  24.5  14.5
## 7 (Paris:~ 2001     80    NA    1.4    6  14.5    0  14.5    0  14.5  14.5
## 8 (T)Raum~ 2004     87    NA    5.6 1275  14.5   4.5   4.5   4.5  14.5  14.5
## 9 *Corpus~ 2002     92    NA    4.9   36  24.5  14.5   4.5   4.5    0   4.5
## 10 ...Lost 2000      5    NA    6.2   13   4.5    0    0   4.5  24.5  34.5
## # i 10,779 more rows
## # 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>

counts = as.vector(averagerate)
xx <-
  barplot(
    averagerate,
    main = "Average Rating of Movie Genres (2000-2005)",
    names = c('Action', 'Animation', 'Comedy', 'Documentary', 'Drama', 'Short', 'Romance'),
    xlab = "Genre",
    ylim = c(0, max(averagerate) + 4),
    cex.axis=0.7, cex.names=0.7,
    col = "light pink")

```

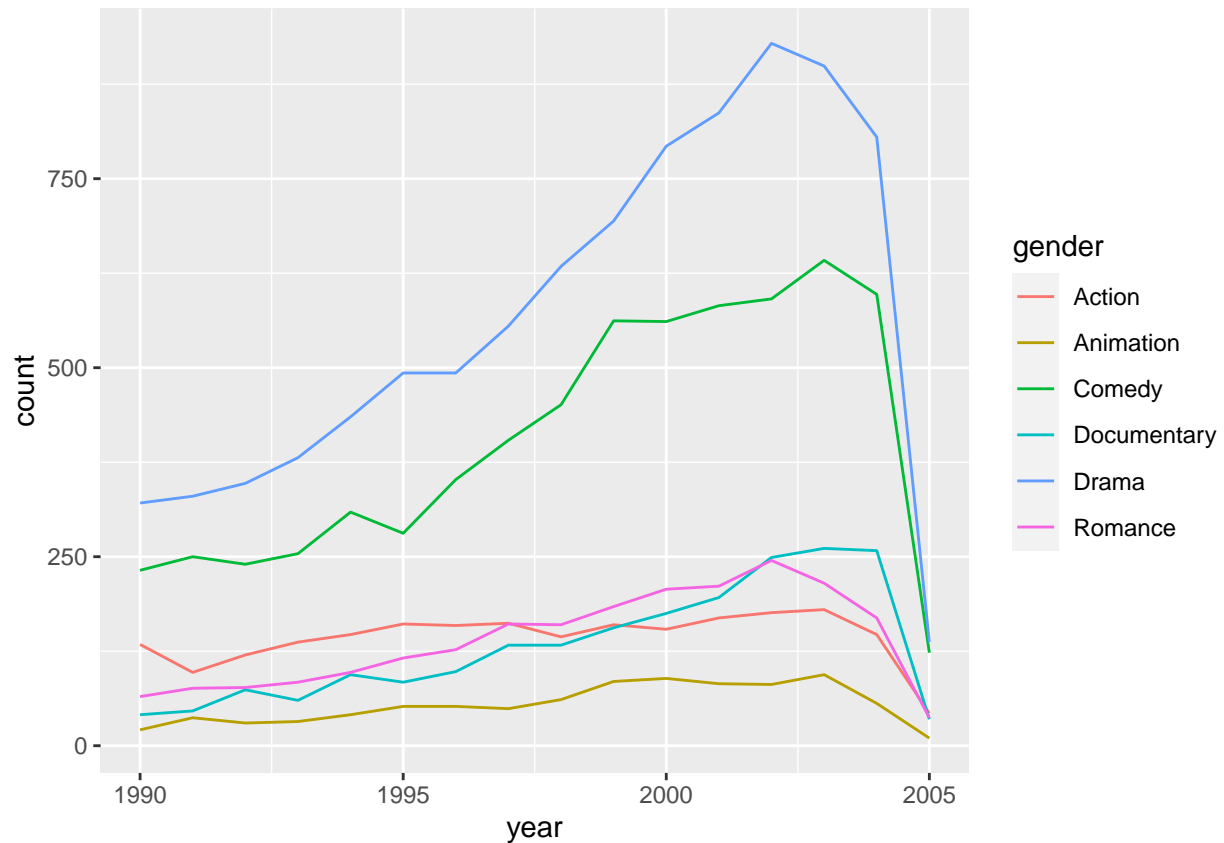
```
text(
  x = xx,
  y = averagerate + 1,
  label = as.character(averagerate),
  cex = 0.7,
  col = "Black")
```

Average Rating of Movie Genres (2000–2005)



Question 8

```
movies %>%
  filter(year > 1989) %>%
  group_by(year) %>%
  summarise_at(vars(Action, Animation, Comedy, Drama, Documentary, Romance), sum) %>%
  pivot_longer(names_to = "gender",
               values_to = "count",
               cols = c(Action, Animation, Comedy, Drama, Documentary, Romance)) %>%
  ggplot() + geom_line(aes(x = year, y = count, color = gender))
```



Question 9

```
# 1st Question: What are the top 5 highest rated movies?
highestrated <- movies %>% arrange(desc(rating)) %>% select(title, rating)
head(highestrated, 5)
```

```
## # A tibble: 5 x 2
##   title                                rating
##   <chr>                                <dbl>
## 1 Dimensia Minds Trilogy: The Hope Factor    10
## 2 Fishing for Love                           10
## 3 Summer Sonata, A                          10
## 4 Black Canyon                             9.9
## 5 Buck Privates                             9.9
```

```
# 2nd Question: What are the top 5 lowest rated movies?
lowestrated <- movies %>% arrange(desc(rating)) %>% select(title, rating)
tail(lowestrated, 5)
```

```
## # A tibble: 5 x 2
##   title                                rating
##   <chr>                                <dbl>
## 1 Vente a ligar al Oeste                1
## 2 Vicious Years, The                    1
## 3 What's Up Front!                     1
## 4 Wild Ones on Wheels                   1
```

5 Windsplitter, The

1

#3rd Question: Where does the median distribution of IMDB movies fall under the rating scale?

```
xx <- hist(movies$rating, main = "Movie Rating Scale",  
  xlab = "Rating Scale",  
  ylab = "# of Movies", col = "light pink")
```

