Homework 5

Abbas Jumani

2023-10-23

```
movies_data <- read.csv("MoviesData.csv")
movies_streaming <- read.csv("MoviesStreaming.csv")
# Load necessary libraries
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(stringr)
```

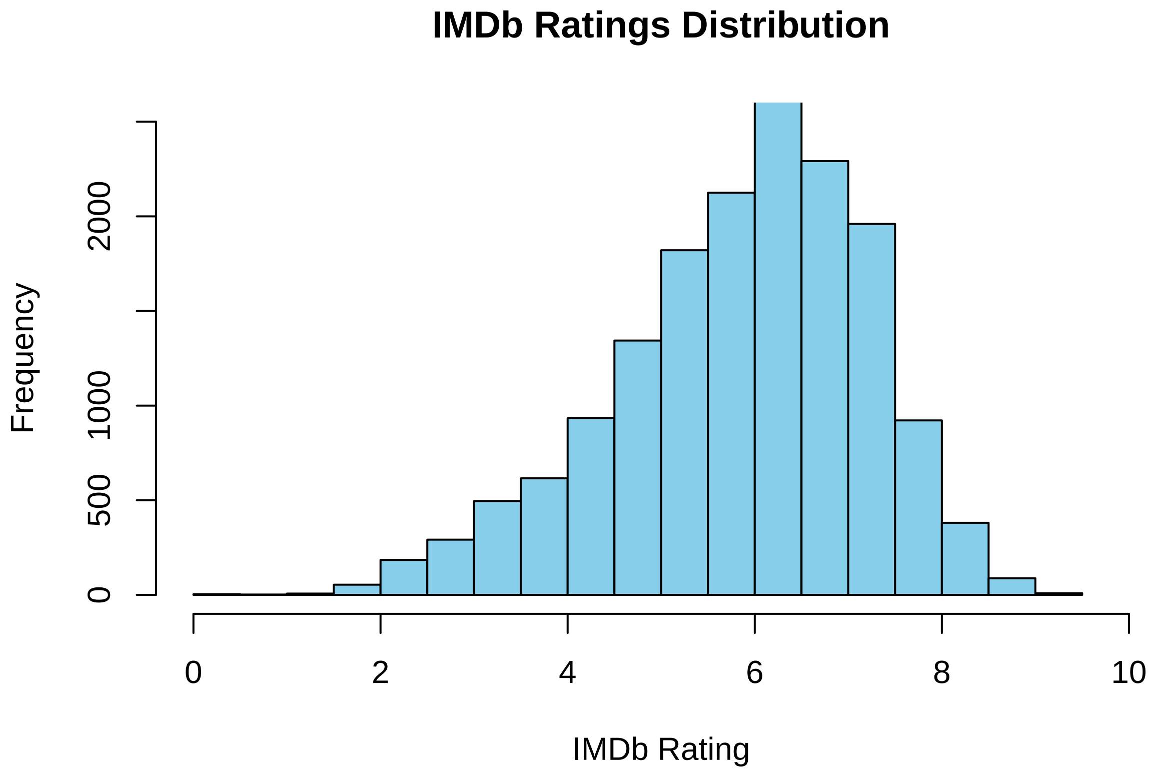
1. Merge two datasets by ID

```
merged_movie_data <- merge(movies_data, movies_streaming, by= "ID")
head(merged_movie_data)
```

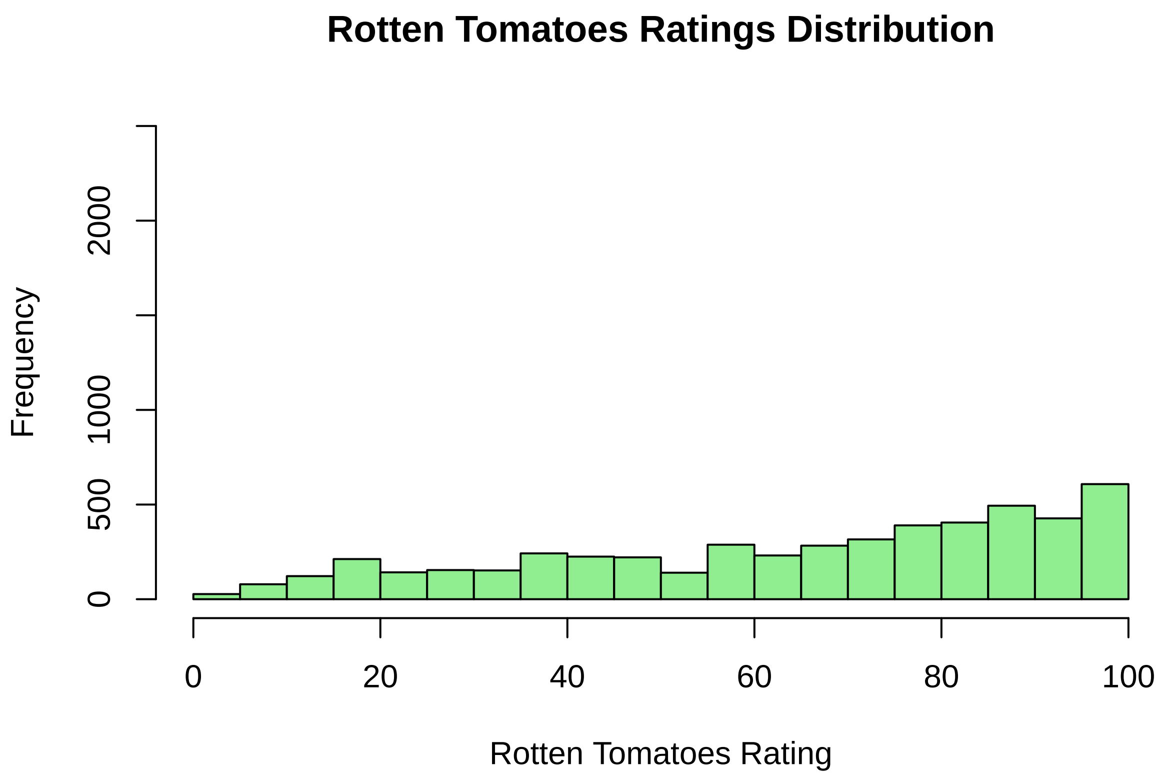
```
##   ID                             Title Year Age IMDb Rotten.Tomatoes
## 1  1                         Inception 2010 13+  8.8              87
## 2  2                        The Matrix 1999 18+  8.7              87
## 3  3            Avengers: Infinity War 2018 13+  8.5              84
## 4  4                Back to the Future 1985  7+  8.5              96
## 5  5    The Good, the Bad and the Ugly 1966 18+  8.8              97
## 6  6 Spider-Man: Into the Spider-Verse 2018  7+  8.4              97
##                                      Directors    Genres Runtime       Country
## 1                            Christopher Nolan    Action     148 United States
## 2               Lana Wachowski,Lilly Wachowski    Action     136 United States
## 3                      Anthony Russo,Joe Russo    Action     149 United States
## 4                              Robert Zemeckis Adventure     116 United States
## 5                                 Sergio Leone   Western     161         Italy
## 6 Bob Persichetti,Peter Ramsey,Rodney Rothman Animation     117 United States
##   Netflix Hulu Prime.Video DisneyPlus
## 1       1    0           0          0
## 2       1    0           0          0
## 3       1    0           0          0
## 4       1    0           0          0
## 5       1    0           1          0
## 6       1    0           0          0
```

#2 Create a histogram for the IMDb ratings, and a separate histogram for theRotten.Tomatoes ratings. Change the labels, the xlim and ylim values (Hint: use rangecommand), and the color on each histogram.

```
hist(merged_movie_data$IMDb,
     main = "IMDb Ratings Distribution",
     xlab = "IMDb Rating",
     ylab = "Frequency",
     col = "skyblue",
     xlim = c(0, 10),
     ylim = c(0, 2500),
     breaks = 20
)
```



```
hist(merged_movie_data$Rotten.Tomatoes,
     main = "Rotten Tomatoes Ratings Distribution",
     xlab = "Rotten Tomatoes Rating",
     ylab = "Frequency",
     col = "lightgreen",
     xlim = c(0, 100),
     ylim = c(0, 2500),
     breaks = 20
)
```



#3 How many movies are in the Adventure Genre?

```
adventure_movies <- merged_movie_data %>%
  filter(Genres == "Adventure")


num_adventure_movies <- nrow(adventure_movies)
```

#4 Create a subset of the data named movie_info that only consists of the followingvariables: ID Title Year, and Runtime. Display the first 6 lines of this new dataframe.

```
movie_info <- merged_movie_data %>%
  select(ID, Title, Year, Runtime)
head(movie_info, n = 6)
```

```
##   ID                        Title Year Runtime
## 1  1                    Inception 2010     148
## 2  2                   The Matrix 1999     136
## 3  3       Avengers: Infinity War 2018     149
## 4  4                 Back to the Future 1985     116
## 5  5    The Good, the Bad and the Ugly 1966     161
## 6  6 Spider-Man: Into the Spider-Verse 2018     117
```

#5 What is the average IMDb rating of all Drama movies? Hint: Use mean () function.

```
# Filter Drama movies and calculate the average IMDb rating
average_imdb_drama <- merged_movie_data %>%
  filter(str_to_lower(Genres) == "drama") %>%
  summarise(average_imdb = mean(IMDb, na.rm = TRUE)) %>%
  pull(average_imdb)

# Print the result
cat("The average IMDb rating of all Drama movies is:", average_imdb_drama, "\n")
```

```
## The average IMDb rating of all Drama movies is: 6.042693
```

#6 What is the highest-rated IMDb Horror from 2015? Your result should only print out themovie title.

```
highest_rated_horror_2015 <- merged_movie_data %>%
  filter(str_to_lower(Genres) == "horror" & Year == 2015) %>%
  top_n(1, wt = IMDb) %>%
  select(Title) %>%
  pull()

# Print the result
cat("The highest-rated IMDb Horror movie from 2015 is:", highest_rated_horror_2015, "\n")
```

```
## The highest-rated IMDb Horror movie from 2015 is: Green Room
```

#7 How many movies are available on Neflix, Hulu, Prime.Video, and Disney Plus? (Forexample, your answer should include: There are _____ movies available on Netflix; Thereare ___ movies available on Hulu. There are _____ movies available on Prime.Video. Thereare _____ movies available on DisneyPlus)

```
# Count movies available on each streaming platform
netflix_count <- sum(merged_movie_data$Netflix > 0, na.rm = TRUE)
hulu_count <- sum(merged_movie_data$Hulu > 0, na.rm = TRUE)
prime_count <- sum(merged_movie_data$Prime.Video > 0, na.rm = TRUE)
disney_count <- sum(merged_movie_data$DisneyPlus > 0, na.rm = TRUE)

# Print the results
cat("There are", netflix_count, "movies available on Netflix;\n")
```

```
## There are 3560 movies available on Netflix;
```

```
cat("There are", hulu_count, "movies available on Hulu;\n")
```

```
## There are 903 movies available on Hulu;
```

```
cat("There are", prime_count, "movies available on Prime Video;\n")
```

```
## There are 12354 movies available on Prime Video;
```

```
cat("There are", disney_count, "movies available on Disney Plus.\n")
```

```
## There are 564 movies available on Disney Plus.
```

#8 How many movies are available via DisneyPlus OR Prime.Videos? (Hint: there are somemovies available on both that should not be double counted in your answer)

```
unique_movies_count <- merged_movie_data %>%
  filter(DisneyPlus > 0 | Prime.Video > 0) %>%
  summarise(unique_count = n_distinct(ID))

# Print the result
cat("There are", unique_movies_count$unique_count, "unique movies available on Disney Plus or Prime Vid
```

```
## There are 12899 unique movies available on Disney Plus or Prime Video.
```