DAS Group Project 2

Group 7

```
library(ggplot2)
library(tidyverse)
library(gt)
library(patchwork)
library(gridExtra)
library(moderndive)
library(GGally)
library(corrplot)
```

```
data <- read.csv("/Users/ziluwang/Documents/GitHub/DAS-Project2-Group7/dataset07.csv", na.
```

1 Introduction

Introduction paragraph

2 Exploratory Data Analysis

```
# Check for missing values
colSums(is.na(data))
```

```
film_id    year  length  budget   votes   genre  rating
      0       0      92       0       0       0       0
```

```
# Data wrangling
data$length[is.na(data$length)] <- median(data$length, na.rm = TRUE)
# Creating a new binary variable
```

```r
data$above_7 <- ifelse(data$rating > 7, 1, 0)
```

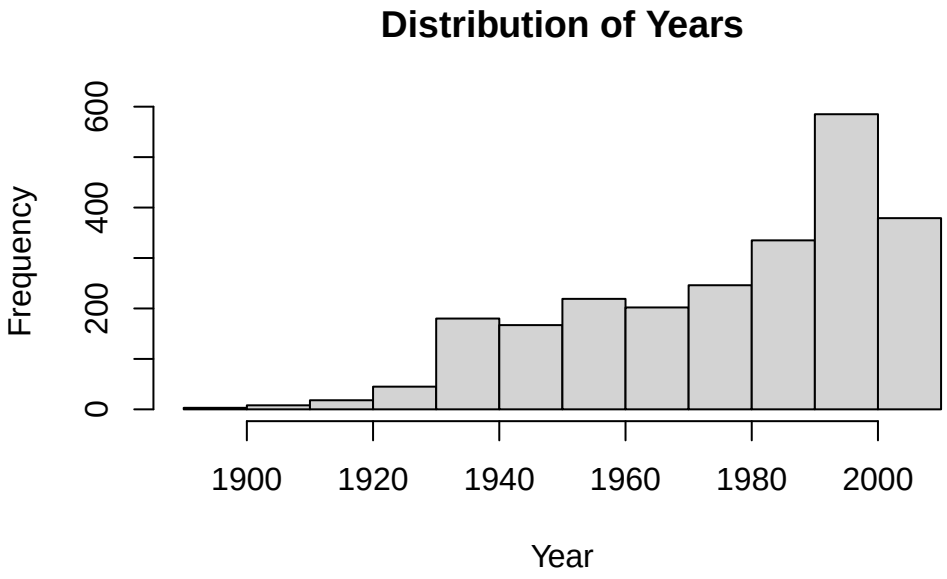
```r
glimpse(data)
```

```
Rows: 2,387
Columns: 8
$ film_id <int> 39891, 33810, 20282, 33131, 50633, 37020, 55337, 28037, 13291,~
$ year    <int> 2003, 2004, 1941, 1959, 1917, 1934, 2003, 1988, 1981, 1935, 19~
$ length  <int> 75, 120, 78, 106, 70, 64, 91, 101, 78, 7, 21, 90, 99, 101, 110~
$ budget  <dbl> 10.9, 19.6, 11.7, 12.0, 14.8, 11.6, 12.6, 10.1, 14.2, 6.6, 10.~
$ votes   <int> 17, 21, 14, 14, 9, 8, 182, 274, 61, 10, 5, 8, 349, 24, 20168, ~
$ genre   <chr> "Action", "Documentary", "Action", "Drama", "Drama", "Drama", ~
$ rating  <dbl> 4.4, 7.3, 2.7, 4.9, 5.6, 4.7, 4.4, 4.3, 4.3, 8.8, 7.3, 8.3, 7.~
$ above_7 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1,~
```

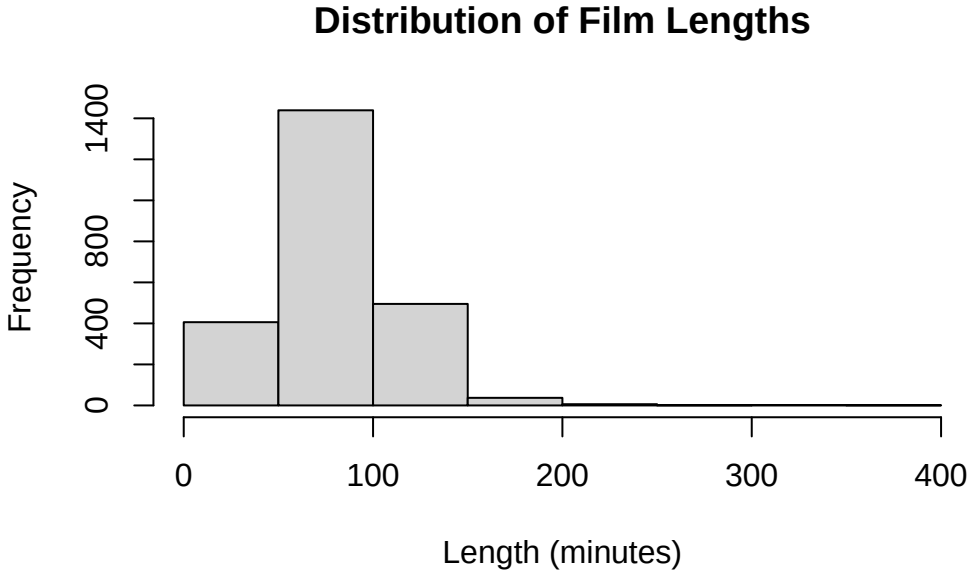
```r
# Summary statistics for each variable
summary_stats <- summary(data)
print(summary_stats)
```

```
    film_id           year          length           budget      
 Min.   :   33   Min.   :1894   Min.   :  1.00   Min.   : 2.10  
 1st Qu.:14799   1st Qu.:1958   1st Qu.: 74.00   1st Qu.:10.00  
 Median :30259   Median :1984   Median : 90.00   Median :12.00  
 Mean   :29942   Mean   :1977   Mean   : 81.75   Mean   :11.95  
 3rd Qu.:44670   3rd Qu.:1998   3rd Qu.:100.00   3rd Qu.:13.90  
 Max.   :58780   Max.   :2005   Max.   :399.00   Max.   :23.70  
     votes           genre               rating         above_7      
 Min.   :     5   Length:2387        Min.   :0.700   Min.   :0.0000  
 1st Qu.:    12   Class :character   1st Qu.:3.700   1st Qu.:0.0000  
 Median :    32   Mode  :character   Median :4.700   Median :0.0000  
 Mean   :   659                      Mean   :5.414   Mean   :0.3523  
 3rd Qu.:   118                      3rd Qu.:7.800   3rd Qu.:1.0000  
 Max.   :103854                      Max.   :9.200   Max.   :1.0000  
```

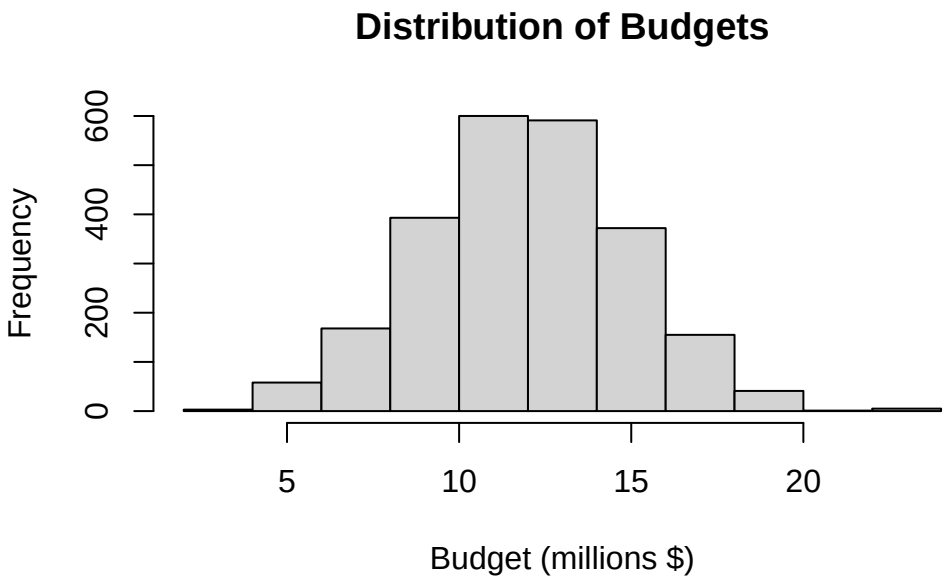
```r
# Check the distribution of numeric variables
hist(data$year, main = "Distribution of Years", xlab = "Year")
```



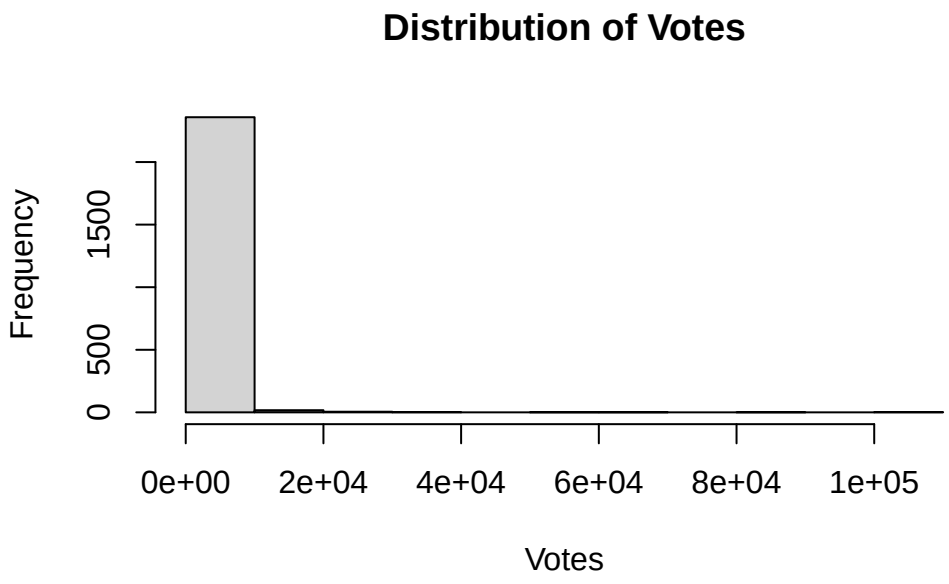

```r
hist(data$length, main = "Distribution of Film Lengths", xlab = "Length (minutes)")
```



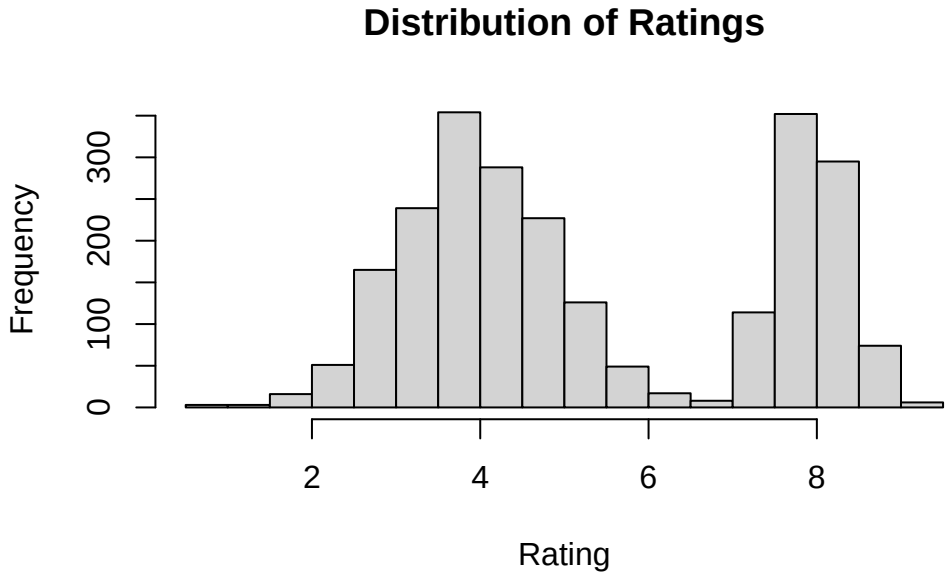
```r
hist(data$budget, main = "Distribution of Budgets", xlab = "Budget (millions $)")
```



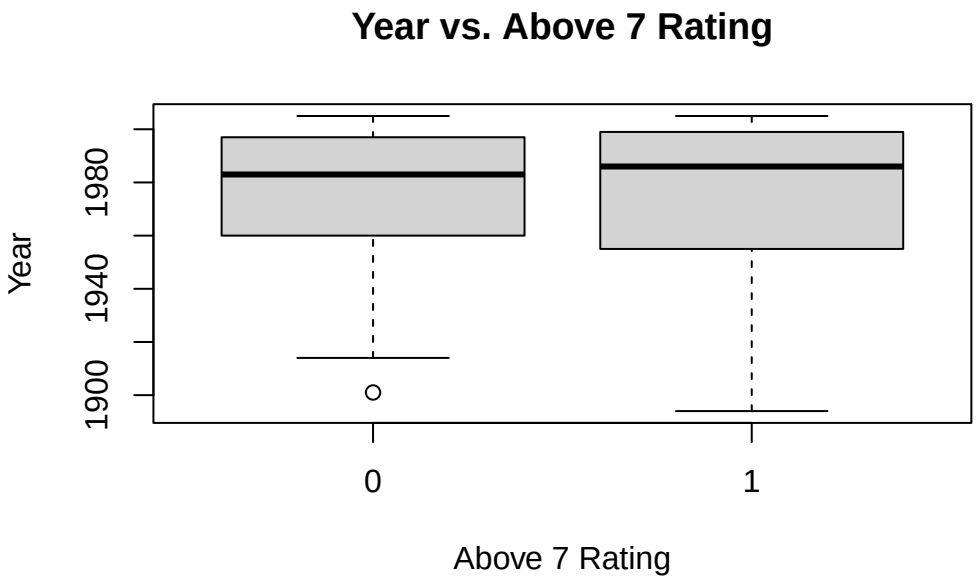

```r
hist(data$votes, main = "Distribution of Votes", xlab = "Votes")
```



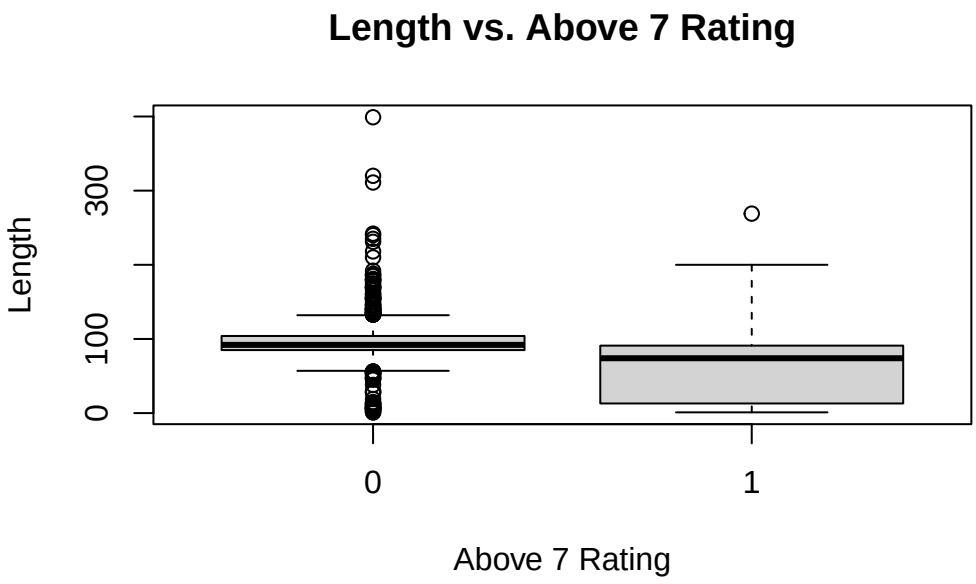

```r
hist(data$rating, main = "Distribution of Ratings", xlab = "Rating")
```



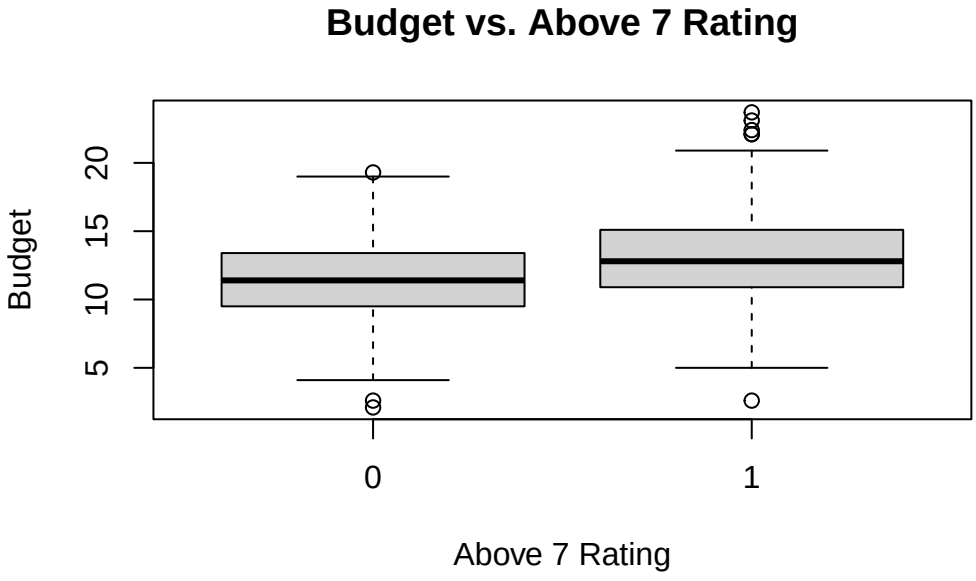
```r
boxplot(data$year ~ data$above_7, main = "Year vs. Above 7 Rating", xlab = "Above 7 Rating
```



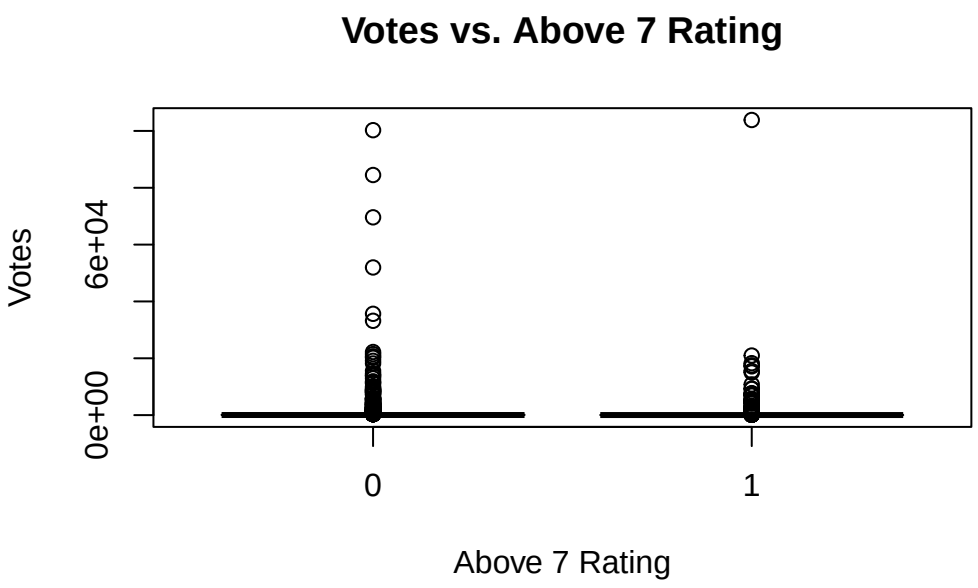

```r
boxplot(data$length ~ data$above_7, main = "Length vs. Above 7 Rating", xlab = "Above 7 Ra
```



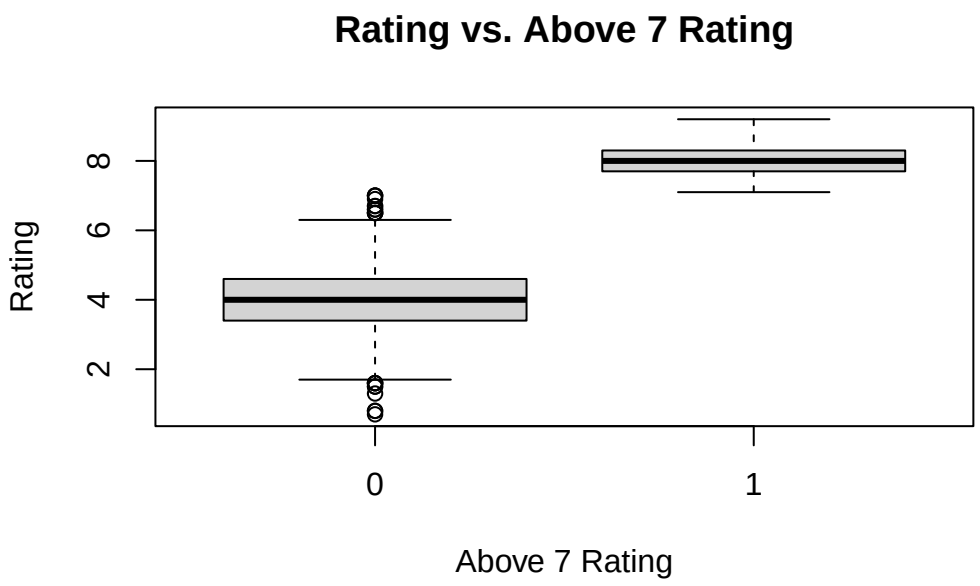

```
boxplot(data$budget ~ data$above_7, main = "Budget vs. Above 7 Rating", xlab = "Above 7 Ra
```



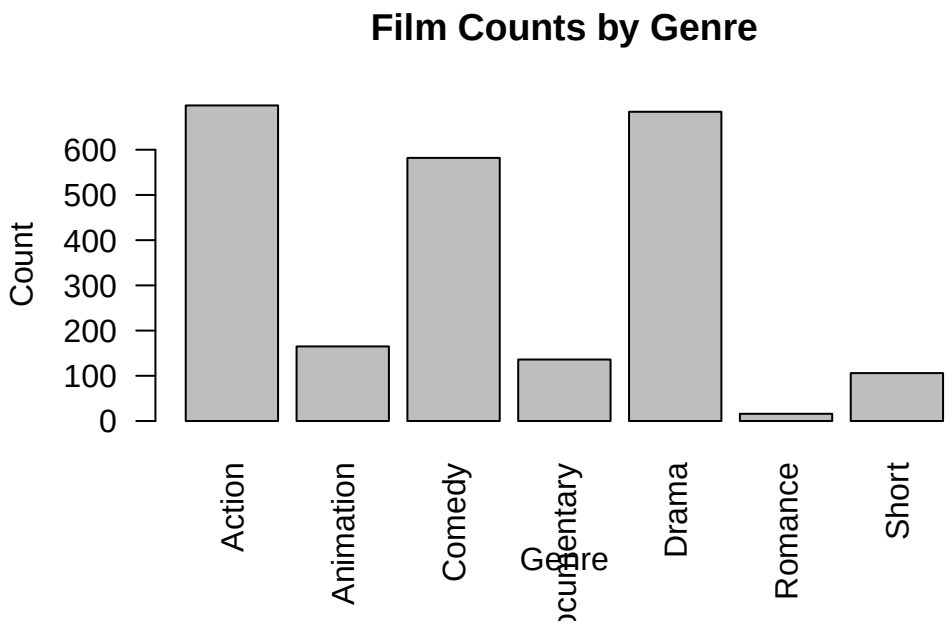
```r
boxplot(data$votes ~ data$above_7, main = "Votes vs. Above 7 Rating", xlab = "Above 7 Rati
```



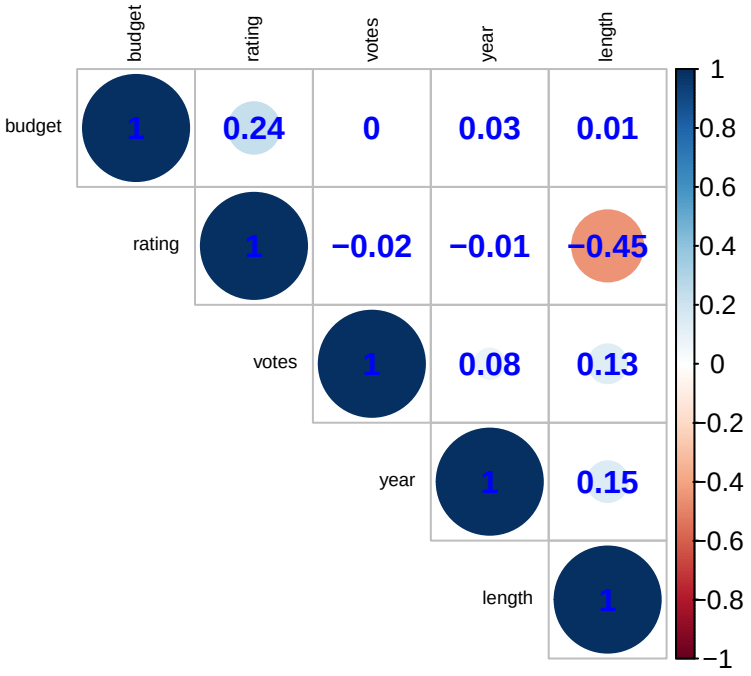

```r
boxplot(data$rating ~ data$above_7, main = "Rating vs. Above 7 Rating", xlab = "Above 7 Ra
```




```r
# Bar plot for genre
genre_counts <- table(data$genre)
barplot(genre_counts, main = "Film Counts by Genre", xlab = "Genre", ylab = "Count", las =
```




```
# Pairwise correlation between numeric variables
numeric_data <- dplyr::select(data, -film_id, -genre, -above_7)  # Remove non-numeric and
cor_matrix <- cor(numeric_data, use = "complete.obs")  # Compute correlation matrix
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
         tl.cex = 0.6, tl.col = "black", addCoef.col = "blue")
```



3 Formal Data Analysis

4 Conclusions

5 Reference