Analysis of Harry Potter Movie Success and Earnings

Nerea de la Torre Veguillas (1669013)

June 18, 2024

Contents

1 Introduction

1.1 What is Harry Potter?

Harry Potter is a globally renowned fantasy series created by the British author J.K. Rowling. The series explains the adventures of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry. The central storyline revolves around Harry's quest to overcome the dark wizard Lord Voldemort, who aims to become immortal and conquer the wizarding world.

The Harry Potter franchise has not only captivated readers and audiences worldwide but has also expanded into a vast multimedia empire, including books, movies, theme parks, and various merchandise. The series has been praised for its imaginative storytelling, complex characters, and its themes of friendship, bravery, and the struggle between good and evil.

1.2 The Harry Potter Movies

The Harry Potter series was adapted into a highly successful film franchise produced by Warner Bros. Pictures. The film series consists of eight movies, each corresponding to the seven books in the series, with the final book being split into two films. Below is a list of the Harry Potter movies in order of their release:

1. **Harry Potter and the Philosopher's Stone** - 2001
2. **Harry Potter and the Chamber of Secrets** - 2002
3. **Harry Potter and the Prisoner of Azkaban** - 2004
4. **Harry Potter and the Goblet of Fire** - 2005
5. **Harry Potter and the Order of the Phoenix** - 2007
6. **Harry Potter and the Half-Blood Prince** - 2009
7. **Harry Potter and the Deathly Hallows – Part 1** - 2010
8. **Harry Potter and the Deathly Hallows – Part 2** - 2011

Each film has been critically acclaimed and commercially successful, contributing significantly to the global popularity and financial success of the Harry Potter franchise.

1.3 Project Objective

The main objective of this project is to analyze the success of the Harry Potter movies, taking into account different aspects such as critics rating and box office earnings. By examining factors such as the release year, runtime, and budget of each film, we aim to identify key determinants of the saga's success.

To achieve this, we will employ both parametric and non-parametric methods. Also a linear regression model will be used to predict the return of investment based on various predictors like budget, run time and release year. Additionally, the bootstrap method will be utilized to estimate the distribution of earnings and to obtain confidence intervals for the mean earnings.

This analysis not only provides insights into the financial performance of one of the most successful film franchises in history but also demonstrates the application of statistical methods in understanding and interpreting real-world data.

2 Description of the Dataset

The dataset used for this analysis consists of information regarding the box office earnings of each Harry Potter movie, along with additional details such as the release year, runtime, budget, and movie title. Below is a summary of the dataset:

Movie_ID	Movie_Title	Release_Year	Runtime	Budget	Box_Office
1	Harry Potter and the Philosopher's Stone	2001	152	125000000	1002000000
2	Harry Potter and the Chamber of Secrets	2002	161	100000000	880300000
3	Harry Potter and the Prisoner of Azkaban	2004	142	130000000	796700000
4	Harry Potter and the Goblet of Fire	2005	157	150000000	896400000
5	Harry Potter and the Order of the Phoenix	2007	138	150000000	942000000
6	Harry Potter and the Half-Blood Prince	2009	153	250000000	943200000
7	Harry Potter and the Deathly Hallows Part 1	2010	146	200000000	976900000
8	Harry Potter and the Deathly Hallows Part 2	2011	130	250000000	1342000000

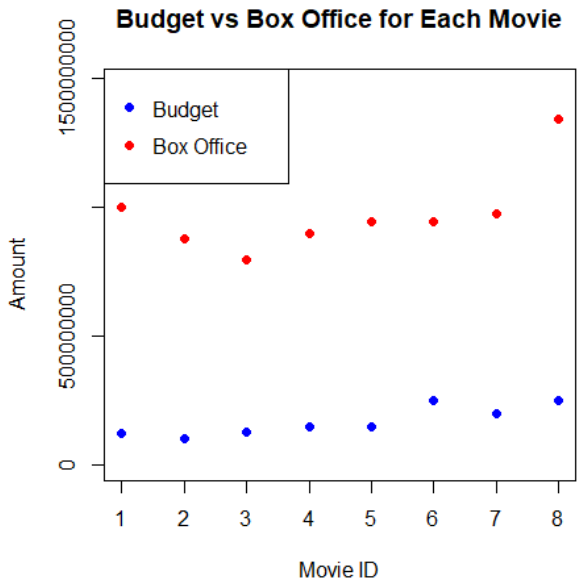
The dataset used for this analysis was obtained from Kaggle, a platform for data science competitions and datasets.

Because our objective is to analyse the success of the films, I thought that it would be very util to add another variable to the dataset: the critics rating. To do so, I searched in IMDb (the world's most popular and authoritative source for movie, TV and celebrity content) for the ratings of the public of every film, and added to the dataset.

Critics_Rating
7.6
7.4
7.9
7.8
7.5
7.6
7.7
8.1

3 Visualization of the data

First, it is util to visualize some characteristics of the dataset. It might be interesting to observe the distribution of Budget and Box Office, because they are two significant variables in the succes of a film.



It is clear that in every film it has been a great profit. We can see that the movie that raised the most money was the last one, but it was also one of the ones that spent the most in budget. We can also observe that the one that spend the least in budget was the second one, but it did not appear to be a disadvantage on the success of the film, due to the high box office. We will study these events in order to identify which may be the most successful film.

Additionally, I calculated the mean of the critics rating, the budget and the box office: 7.7, 16.9375.000, 97.2437.500 respectively.

4 Analysis of the data

4.1 Correlation of variables

Before starting analysing the success and the profits of the saga, we shall observe if there are any relation between the variables. I found interesting to compare the next ones:

Critics Rating - Box Office: 0.5698311
Budget - Box Office: 0.6390482
Release Year - Critics Rating: 0.5418672

We know that a value close to 1 indicates a strong positive correlation. In our case, this means that as box office revenues increase, critics' ratings also tend to increase, and interestingly, they also increase over the years. This could be due to the experience gradually acquired throughout the saga, improvements in special effects, progressive improvements in the script and cinematography...
We can also observe that the most positively related variables are budget and earnings.

4.2 Non-parametric bootstrap

The non-parametric bootstrap involves using the frequency distribution of the n data points as an estimate of the population or probability distribution. Then, it simulates sampling from this estimated population distribution to obtain new bootstrap samples. For each sample, some statistic of interest can be calculated.

In our case, may be interesting to analyse which film is the most "successful".

4.2.1 Estimation of the most profitable film

A good way of analysing success is the return on investment, in other words, how profitable the movie was.
To do this, we have to take into account the box office earnings and the budget. Using non-parametric bootstrap we can estimate confidence in the most profitable movie, based on the variability of bootstrap samples.

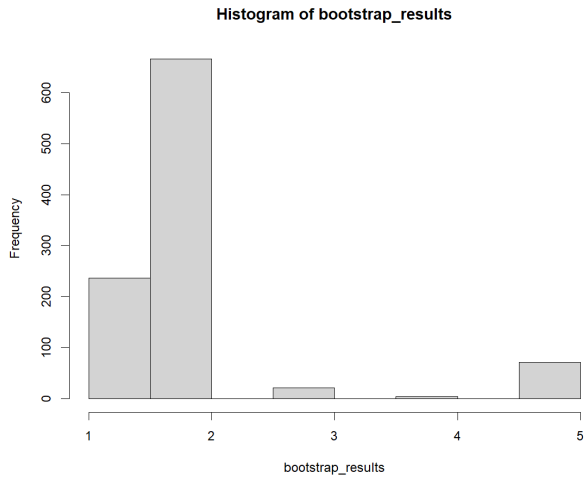
To do so, I decided to add a new column to the dataset containing the ratio of box office revenue to budget (ROI). Next, we will generate 1000 random samples, with the 8 movies in each, using the data collected.

Finally we will calculate how frequently each movie is the most profitable in the bootstrapped samples, and identify the movie that appears most frequently as the most "successfull".

Once implemented in the code (see last section). The frequencies for each film are:

1	2	3	4	5	6	7	8
237	667	21	04	71	0	0	0

The graphic generated to analyse the data is the following:



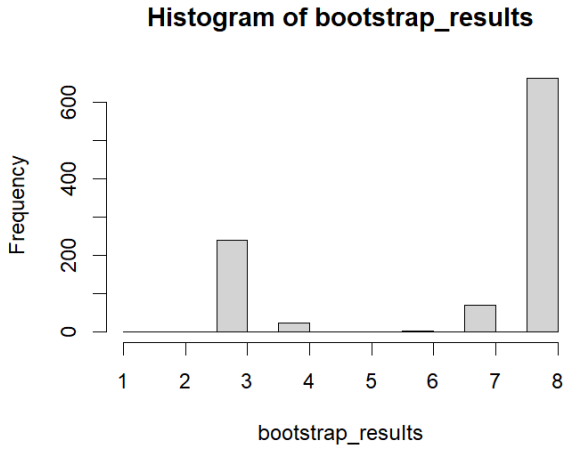
We can observe that in the histogram generated the highest frequencies are, as in the table, the second film. We can also see that the last movies, 6, 7 and 8 do not even appear. The explanation of this is that, if we take a look into the budget's column of the dataset, we can observe that there is huge difference of budget between the last 3 movies and the others (lots of more millions). Consequently, eventhough the box office earinings are high, (specially the last one) the ROI is lower in comparison with the others.

With all the information and the graphics above, we can confirm that the most "successful" movie of the saga, in terms of budget and earnings, is the second one, Harry Potter and the Chamber of Secrets.

4.2.2 Estimation of the most liked film

We have analysed the success of the films with the return on investment, but success is much more than money. We should take into account also the opinion of the people.
So I decided to implement the same analysis, but now we will calculate how frequently each movie is the most "liked" in the bootstrapped samples, looking at the value of the critics rating.

The results were:



We can see that the movie with more frequencies of being the most liked is the last one, Harry Potter and the Deathly Hallows part 2. This makes sense because the highest value of critics rating is to this film.

4.2.3 Estimation of the most successful film

We have been able to see which movies are the most successful in terms of economic success and audience success. Although we have made an accurate analysis to make sure, these are aspects that can be easily interpreted by simply analyzing the data from our original dataset, such as the critics' rating. Therefore, it would be ideal to go a bit further and understand success as a combination of different aspects, rather than just black or white. For this reason, I believe the best way to analyze which movie is the most successful is by considering both factors simultaneously.

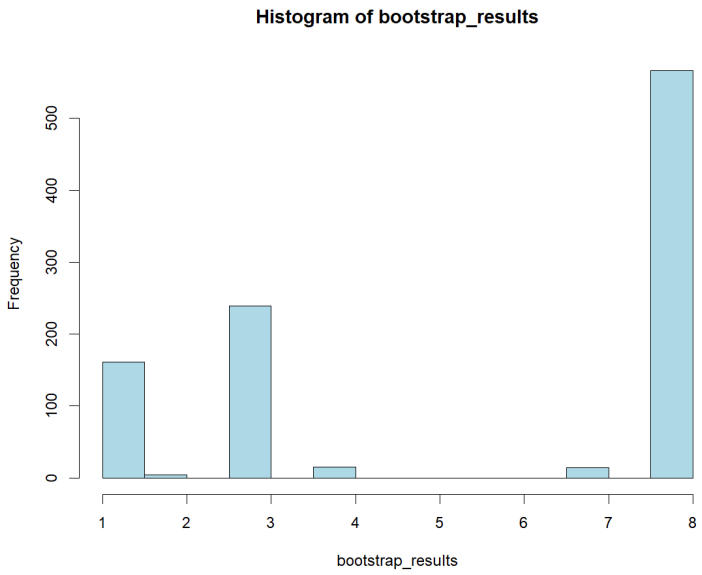
To conduct a combined analysis that considers both profitability (ROI) and critical acclaim (Critics Rating) and select the movie with the best values in both aspects, I performed a non-parametric bootstrap that encompasses both factors.
First, it is necessary to calculate a combined score. Specifically, I used a weighted combination of ROI and Critics Rating by averaging both normalized values. This new score is added to the sample dataset to subsequently identify the movie with the highest value.

Finally, I evaluate the frequency with which each movie appears as the best in both metrics. (we are still generating 1000 random samples)

The frequencies for each film are:

1	2	3	4	5	6	7	8
161	4	239	15	0	0	14	567

The graphic generated to analyse the data is the following:



We can see that even though the second film had the best return on investment, the fact that it is one of the least liked by the public makes it barely appear as the most successful. On the other hand, the third film, which is one of the most loved by the public and has a relatively high ROI, has increased its frequency compared to the separate analyses. Finally we see that there is a clear winner, the last film, due to its box office success and being the most liked by the public.

4.2.4 Estimation of the ROI

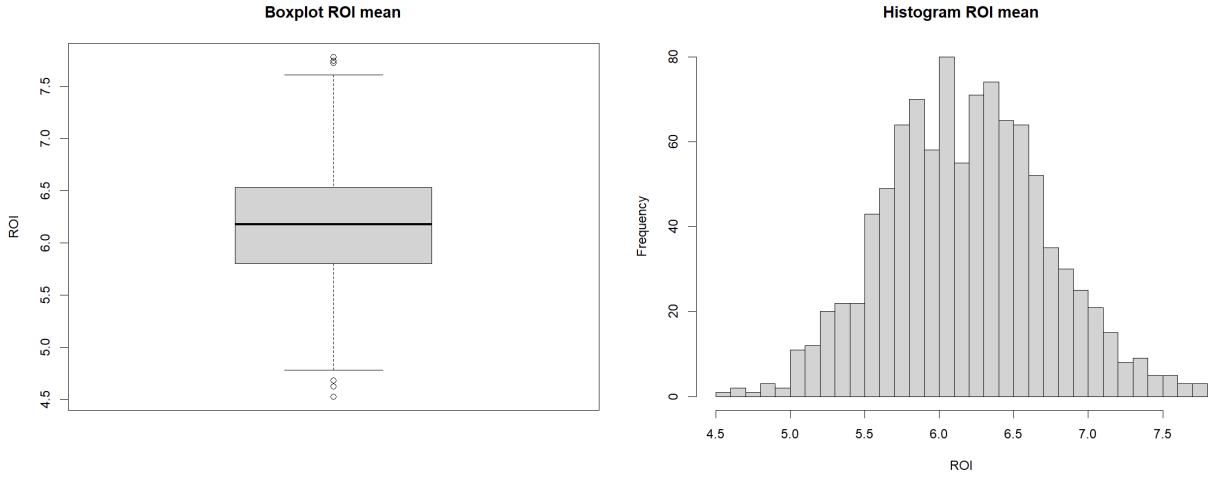
Finally, in this section, we focus on evaluating the profitability (ROI) of each movie in the Harry Potter saga. To achieve this, we used the non-parametric bootstrap method, using 1000 samples of the 8 films, to estimate the average profitability, its standard deviation, and the 95% confidence interval.

A fraction of the output is:

5.841325 6.644865 5.916258 5.954745 7.327783 5.270825 5.917653 5.937778 6.149263 6.446533 5.762273

The mean of the ROI of all the samples is: 6.175894.
The standard error is: 0.540932.
The confidence interval of 95% is: [5.185607 - 7.291713]

The graphics we created to analyse the data are the following:



The box plot shows that there are a few outliers in the mean, these are caused by the high values of the first and second movies, and the lower value of the sixth one in the data set. On the other hand, the mean of all the datasets (the line in black) and the majority of the means land between 5.75 and 6.5.
The histogram generated shows that the highest frequencies are, as in the box plot, the ones between 5.75 and 6.5. With the information above, we see that the standard error, 0.540932,

is small, indicating that the results are close in value to the mean. With all the data and graphics above, and also with the confidence interval [5.185607 - 7.291713], we can say that the mean of the return on investment of the saga is approximately 6.175894.

4.3 Parametric bootstrap

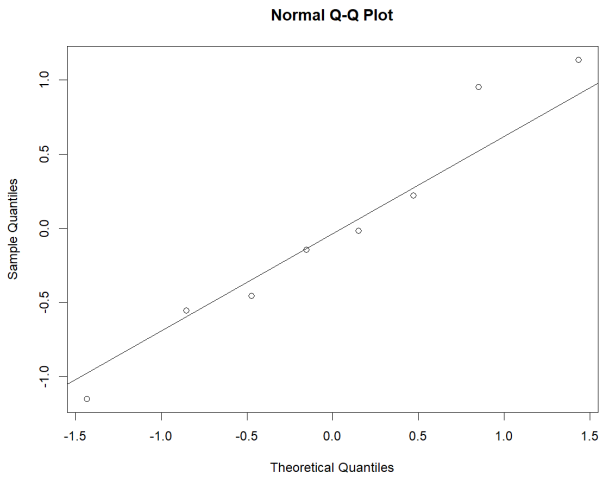
Parametric bootstrapping operates under the assumption that the data is derived from a specific distribution with unknown parameters.
We assume that the data originates from random variables following a particular distribution, such as Normal, Poisson...
Then we use the data to estimate the parameters of this distribution (mean, variance...) and we generate samples using a random number generator based on the estimated distribution, ensuring these samples are of the same size as the original dataset.
Finally we can compute the sample mean or other statistics of interest.

4.3.1 ROI Prediction Using a Regression Model

The objective is to use the original data to fit a regression model to predict ROI based on variables such as critics rating, duration, and budget, and use parametric bootstrap to assess the stability of the coefficients. This model, can help us to estimate the return of investment of the Harry Potter films if we own basic data of each film, like the year, the runtime or the budget.

To do so, I fitted a linear regression model and used parametric bootstrap to generate samples from the residuals' distribution.

First of all, in order to use parametric bootstrap, we have to assume that our dataset follows an specific distribution. In this case, I will assume that the standard errors of the dataset follow a Normal distribution. But to make sure of it, I used the Q-Q (Quantile-Quantile) plot, which is a visual tool used to evaluate if a data sample follows a specific theoretical distribution, such as the normal distribution in this case. The goal is to compare the observed quantiles of the data with the expected quantiles of a theoretical distribution (in this case, the normal distribution).
So, we fit the regression model to predict ROI based on release year, runtime, and budget. Then, we create the Q-Q plot of the model's residuals. The output of the graphic is:



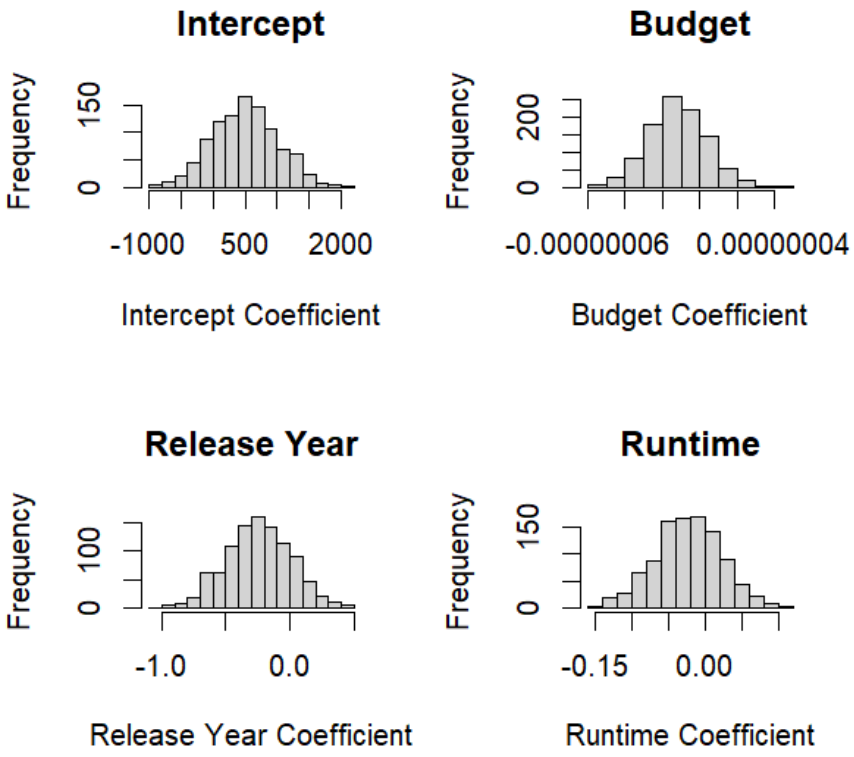
We can see that the points in the Q-Q plot approximately align along a diagonal line (the Q-Q line), which indicates that the data roughly follows a normal distribution.

On the other hand, the regression model coefficients are as follows:
(Intercept) = 475.35984296834
Budget = -0.00000001277
Release Year = -0.23104766208
Runtime = -0.02396861553

Now we will run the bootstrap to see the distribution of the model coefficients, which will give us an idea of their variability and the ROI.

Once we know this, we can set up the necessary parameters for the bootstrap, including the number of observations, and the vectors to store the coefficients. In each iteration, we generate random errors from a normal distribution with mean 0 and standard deviation equal to that of the original model residuals. We generate new bootstrap data using the original model coefficients and the random errors. We fit a regression model to the bootstrap data and store the coefficients of the fitted model.

To see the distribution of the values of the different coefficients, I have created the following graphs:



We can see that the 4 coefficients follow an approximately normal distribution and that the mean coincides with the coefficients of the initial regression model, as we expected.

Finally, to verify that the model predicts well, I first calculated the mean of the ROI values from the bootstrap samples, which was: mean(rROI) = 6.449728. We can see that this closely approximates the mean value calculated using non-parametric bootstrap.

Additionally, I predicted the ROI for the movie "Harry Potter and the Order of the Phoenix" (using its respective data) to verify that the prediction is accurate. The result was 6.423606, which is very close to the original value of 6.28.

5 Conclusions

After conducting the initial analysis using non-parametric bootstrap to estimate the success of the Harry Potter movies, it was found that "Harry Potter and the Deathly Hallows Part 2" was identified as the most successful movie most frequently in the bootstrap samples, due to its higher values on the box office and the critics rating. We also observed that although "Harry Potter and the Chamber of Secrets" has the highest return of investment by far, its low critics rating means it does not frequently appear as the most successful movie.
We can also conclude that the distribution of the estimated profitability showed significant variability among the movies, highlighting differences in the expected return on investment (ROI). These results suggest that while some movies may have higher box office revenues, the profitability adjusted for the invested budget can vary considerably.
Furthermore, with our fourth analysis using the non-parametric bootstrap method, we demonstrated that the Harry Potter movies have had consistently high average profitability, although with some variability between the movies.

Regarding the analysis with parametric bootstrap, we can say that, although the sample size is small, we have been able to analyze the evaluation of ROI and how it can be affected by variables such as release year, runtime, and budget. However, it is important to interpret the results with caution and consider the limitations inherent in working with such a small dataset. This has been particularly useful for understanding the success of the Harry Potter movies in terms of ROI, release year, and runtime. These findings, also, provide a deeper understanding of the factors that have contributed to the financial success of the franchise.

Finally, the combined use of non-parametric and parametric bootstrap in analyzing the Harry Potter movie dataset has provided us with a comprehensive evaluation of profitability and success, allowing us to identify the movies that stand out the most in these two aspects. The results obtained help to better understand the variability and uncertainty associated with key metrics of the financial performance of the movies.

6 Bibliography

References

[1] IMDb. *Harry Potter Series*. Link: https://www.imdb.com

[2] Parametric Bootstrap presentation from UAB: https://e-aules.uab.cat/2023-24/pluginfile.php/710856/mod_resource/content/1/5-%20Parametric%20Bootstrap_2024.pdf

[3] Non-Parametric Bootstrap presentation from UAB: https://e-aules.uab.cat/2023-24/pluginfile.php/715643/mod_resource/content/1/6-%20Non-Parametric%20Bootstrap.pdf

[4] Dataset from Kaggle: https://www.kaggle.com/datasets/maricinnamon/harry-potter-movies-dataset

7 R scripts

7.1 Visualization of the data

```r
library("readxl")
dataset = read_excel("C:/Users/nerea/Downloads/Movies.xlsx")
dataset$Critics_Rating <- c(7.6, 7.4, 7.9, 7.7, 7.5, 7.6, 7.8, 8.1)

summary(dataset)

boxplot(dataset$Budget, main="Boxplot of Budget")
boxplot(dataset$Box_Office, main="Boxplot of Box_Office")

plot(dataset$Movie_ID, dataset$Budget,
     col = "blue", pch = 16,
     xlab = "Movie ID", ylab = "Amount",
     main = "Budget vs Box Office for Each Movie",
     ylim = c(0, max(dataset$Budget, dataset$Box_Office) * 1.1))
points(dataset$Movie_ID, dataset$Box_Office, col = "red", pch = 16)
legend("topleft", legend = c("Budget", "Box Office"),
       col = c("blue", "red"), pch = 16)

mean(dataset$Critics_Rating)
mean(dataset$Budget)
mean(dataset$Box_Office)
```

7.2 Correlation of variables

```r
cor(dataset$Critics_Rating, dataset$Box_Office)
cor(dataset$Budget, dataset$Box_Office)
cor(dataset$Release_Year, dataset$Critics_Rating)
```

7.3 Non-parametric bootstrap

7.3.1 Estimation of the most profitable film

```r
dataset$ROI <- dataset$Box_Office / dataset$Budget

most_profitable_non_param <- function(data) {
  sample_indices <- sample(1:nrow(data), size = nrow(data), replace =
     TRUE)
  sample_data <- data[sample_indices, ]
  most_profitable <- sample_indices[which.max(sample_data$ROI)]
  return(most_profitable)
}

R <- 1000
set.seed(123)
bootstrap_results_profitable <- replicate(R,
   most_profitable_non_param(dataset))
```

```r

freq_profitable <- table(bootstrap_results_profitable)

most_frequent_profitable_movie <- names(which.max(freq))

hist(bootstrap_results_profitable)
```

7.3.2 Estimation of the most liked film

```r
most_liked_non_param <- function(data) {
  sample_indices <- sample(1:nrow(data), size = nrow(data), replace =
     TRUE)
  sample_data <- data[sample_indices, ]
  most_liked <- sample_indices[which.max(sample_data$Critics_Rating)]
  return(most_liked)
}

R <- 1000
set.seed(123)
bootstrap_results_liked <- replicate(R, most_liked_non_param(dataset))

freq_liked <- table(bootstrap_results)

print(freq_liked)

most_frequent_liked_movie <- names(which.max(freq))

hist(bootstrap_results_liked)
```

7.3.3 Estimation of the most successful film

```r

library(dplyr)

calculate_combined_score <- function(data) {
  data <- data %>%
    mutate(Normalized_ROI = (ROI - min(ROI)) / (max(ROI) - min(ROI)),
           Normalized_Critics_Rating = (Critics_Rating -
              min(Critics_Rating)) / (max(Critics_Rating) -
              min(Critics_Rating)),
           Combined_Score = Normalized_ROI + Normalized_Critics_Rating)
  return(data)
}

  most_profitable_and_liked_non_param <- function(data) {
  sample_indices <- sample(1:nrow(data), size = nrow(data), replace =
     TRUE)
  sample_data <- data[sample_indices, ]

  sample_data <- calculate_combined_score(sample_data)
  best_movie <- which.max(sample_data$Combined_Score)
```

```r

  return(sample_data$Movie_ID[best_movie])
}

R <- 1000
set.seed(123)
bootstrap_results <- replicate(R,
    most_profitable_and_liked_non_param(dataset))

freq <- table(bootstrap_results)

most_frequent_profitable_and_liked_movie <- names(which.max(freq))
hist(bootstrap_results,  col = "lightblue")
```

7.3.4 Estimation of the ROI

```r
non_param_roi <- function(x) {
  x <- mean(sample(x, size = length(x), replace = TRUE))
  return(x)
}

stats_roi <- replicate(1000, non_param_roi(dataset$ROI))
stats_roi

final_roi <- mean(stats_roi)
sd_roi <- sd(stats_roi)
CIroi <- quantile(stats_roi, probs = c(0.025, 0.975))

boxplot(stats_roi, main = "Boxplot ROI mean", ylab = "ROI")
hist(stats_roi, main = "Histogram ROI mean", xlab = "ROI", breaks = 30)
```

7.4 Parametric bootstrap

7.4.1 ROI Prediction Using a Regression Mode

```r
model <- lm(ROI ~ Budget + Release_Year + Runtime, data = dataset)
summary_model <- summary(model)
summary_model

residuals <- residuals(model)
qqnorm(residuals)
qqline(residuals)

nsim <- 1000
n <- nrow(dataset)

intercepts <- numeric(nsim)
slopes_budget <- numeric(nsim)
slopes_release_year <- numeric(nsim)
slopes_runtime <- numeric(nsim)
rROI <- numeric(n)
```

```r

for (i in 1:nsim) {
  error <- rnorm(n, 0, summary_model$sigma)
  rROI <- model$coefficients[1] +
    model$coefficients[2] * dataset$Budget +
    model$coefficients[3] * dataset$Release_Year +
    model$coefficients[4] * dataset$Runtime +
    error
  bootstrap_model <- lm(rROI ~ Budget + Release_Year + Runtime, data =
     dataset)
  intercepts[i] <- bootstrap_model$coefficients[1]
  slopes_budget[i] <- bootstrap_model$coefficients[2]
  slopes_release_year[i] <- bootstrap_model$coefficients[3]
  slopes_runtime[i] <- bootstrap_model$coefficients[4]
}

par(mfrow = c(2, 2))
hist(intercepts, main = "Intercept", xlab = "Intercept Coefficient")
hist(slopes_budget, main = "Budget", xlab = "Budget Coefficient")
hist(slopes_release_year, main = "Release Year", xlab = "Release Year
   Coefficient")
hist(slopes_runtime, main = "Runtime", xlab = "Runtime Coefficient")

mean_bootstrap_ROI <- mean(rROI)
mean_bootstrap_ROI

roi <- model$coefficients[1] +
  model$coefficients[2] * 150000000 +
  model$coefficients[3] * 2007 +
  model$coefficients[4] * 138
roi
```