How film length and budget are positively related to a film's reception, while year of release is negatively related to a film's reception

Arjun Dhatt

December 22, 2020

Keywords:

- Film
- Year
- Budget
- Linear Regression
- Rating

Abstract

In this report, I will be looking at the various factors that influence a film's reception among audiences. Using the 'ggplot2movies' packages, I will be looking at the dataset 'movies' and filtering the dataset to include relevant variables such as — title, rating, length, year, and budget. I will then run an analysis to identify whether the variables 'budget', 'length', or 'year' affect a film's reception. After completing the analysis, I concluded that a film's length and budget have a positive relationship with the film's rating, while year has a negative relationship.

Introduction

Many variables can influence an audience's reception to a film. These variables include— the budget of the film, the length of the film, and the year the film was released. This poses the interesting question — how does a film's budget, length and year of release affect it's reception among audiences?

In the report, I will investigate this question by taking a deep look at the interplay between all the factors. I began the analysis by cleaning the 'ggplot2movies' data so that it only includes variables of interest such as — title, length, rating, budget, and year of release. Since some of these entries in our dataset do not contain information for relevant variables, I deleted those entries so that all films in the dataset are analyzed from the same standpoint.

I began the analysis by looking at the variables budget and length. I conducted linear regression for both variables and discovered that there is a significant positive relationship due to the significantly small p-value. Lastly, I looked at the variable year and discovered that there is a significant negative relationship due to the significantly small p-value. With these results, I concluded that a film's budget and length positively affect a film's reception, while a film's year of release negatively affect a film's reception.

The results from this analysis are important because they provide an outlook on important variables that affect a film's reception. This information can be useful to both audiences that watch films as well as to film production companies so they can forecast what the reception for their films will be like.

In this report, I will begin by explaining the data that I will be using in the 'Data' section, followed by the model I will use to make conclusions in the 'Model' section. I will then show the results from my analysis in

the 'Results' section and explain them in the 'Discussion' sections. Lastly, I will reserve a section for any extranneous information, code, and references in the 'Appendix/References' section.

Data

In this report, I will be using the 'ggplot2movies' dataset. The data is collected from the website https://www.imdb.com — a website dedicated to collecting information about various films, television shows, and celebrities. Users are able to complete reviews on the IMDB website allowing them to provide a score for a film on a scale from 1-10.

The 'ggplot2movies' package selects films from the IMDB site if they had a known length and had been rated by at least one IMDB user. Each film corresponds to one entry in the dataset and contains information on up to 25 variables, but not all films contain information on all 25 categories. For example, many of the films from before 1950 do not contain information about the film budget which is important when determining whether budget affects a films reception. The 'ggplot2movies' package contains information on a total of 58,788 films. All of the films used in the dataset were collected from the year 1893 to 2005.

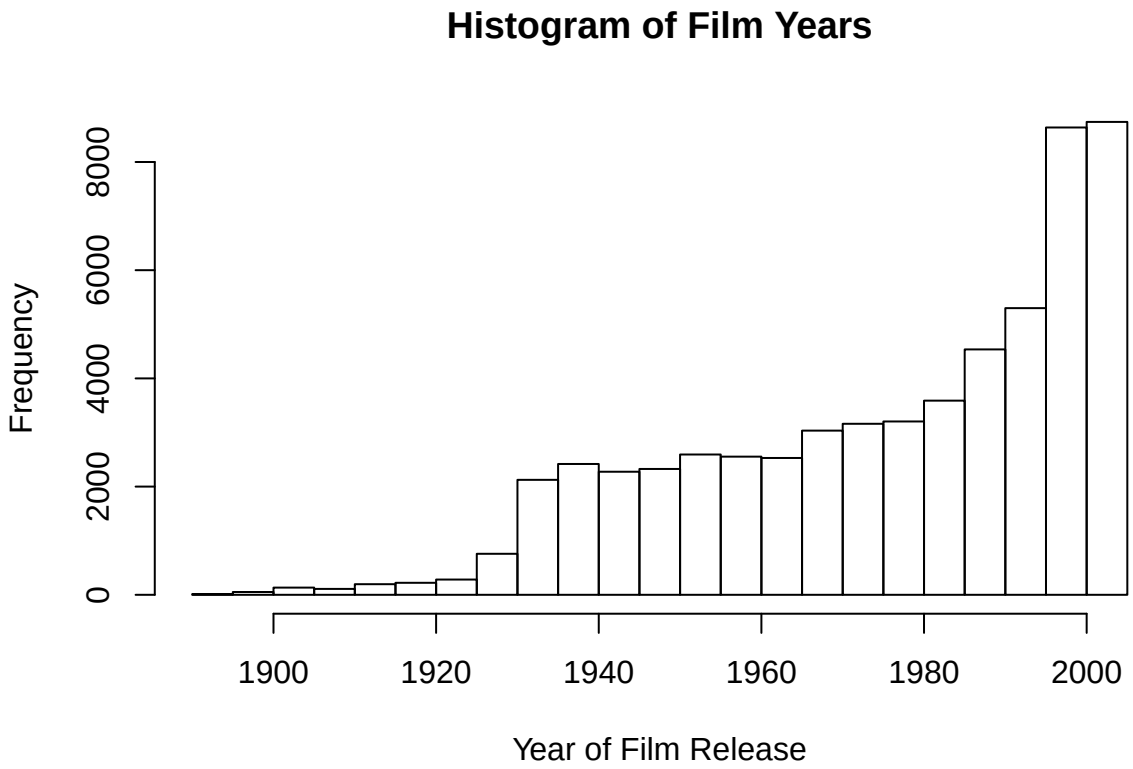
Although the dataset contains information on 25 variables, I will only be analyzing a select few variables in the dataset. The variables and descriptions of the variables I will be analyzing can be seen below:

- 'title' — provides the titles of the movie
- 'year' — provides the year that the film was released
- 'budget' — provides the total budget of the film in US dollars
- 'length' — provides the length of the film
- 'rating' — provides the average rating of the film by IMDB users

A large part of this report will use rating as a dependant variable and rating is largely dependant on the number of votes. The original dataset contains 58,788 films, but many of those films contain very little votes. According to the law of large number, as the sample size grows larger, the rating will approach its true mean; the rating can not approach the true mean if it is based on only a few number of votes. Therefore, I removed all entries with less than 30 votes. The 30 vote threshold was based on the 50th percentile of 'votes'. Removing all entries below the 50th percentile ensures that the rating is not unreliable due to a few number of votes.

When observing a histogram based on year of release, we can see that the dataset is severely left-skewed:

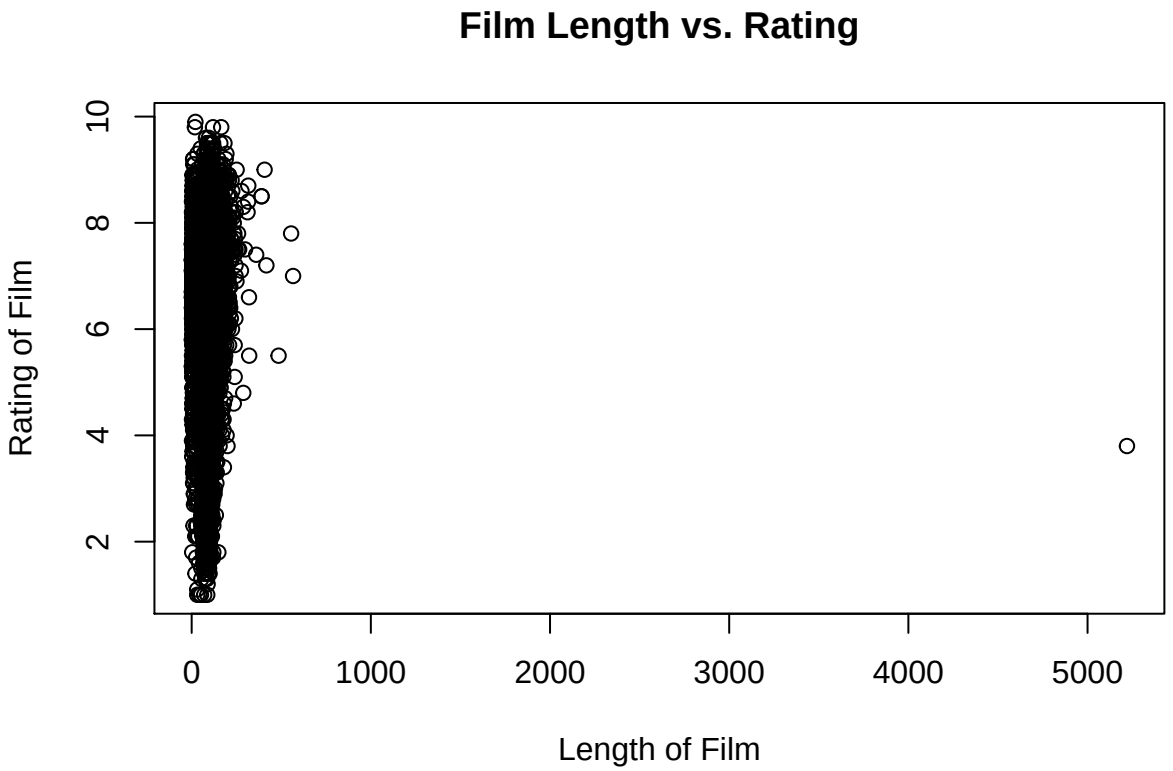
Figure 1



Despite the fact that it is left skewed, there is still a large number of films that are released before 1950; the dataset actually contains 10,418 entries from films before 1950. Films from before 1950 are different from the films that we are used to seeing in today's age in many different ways. Reviewers may have focused on stylistic traits, rather than both stylistic and narrative traits like today's movies; movies were also much shorter. Due to the unfair critical comparisons between the films from different eras, I decided to focus exclusively on films released after 1950's as they conform to the style we are used to seeing in modern films. Therefore, I removed all films from the year 1950 and below.

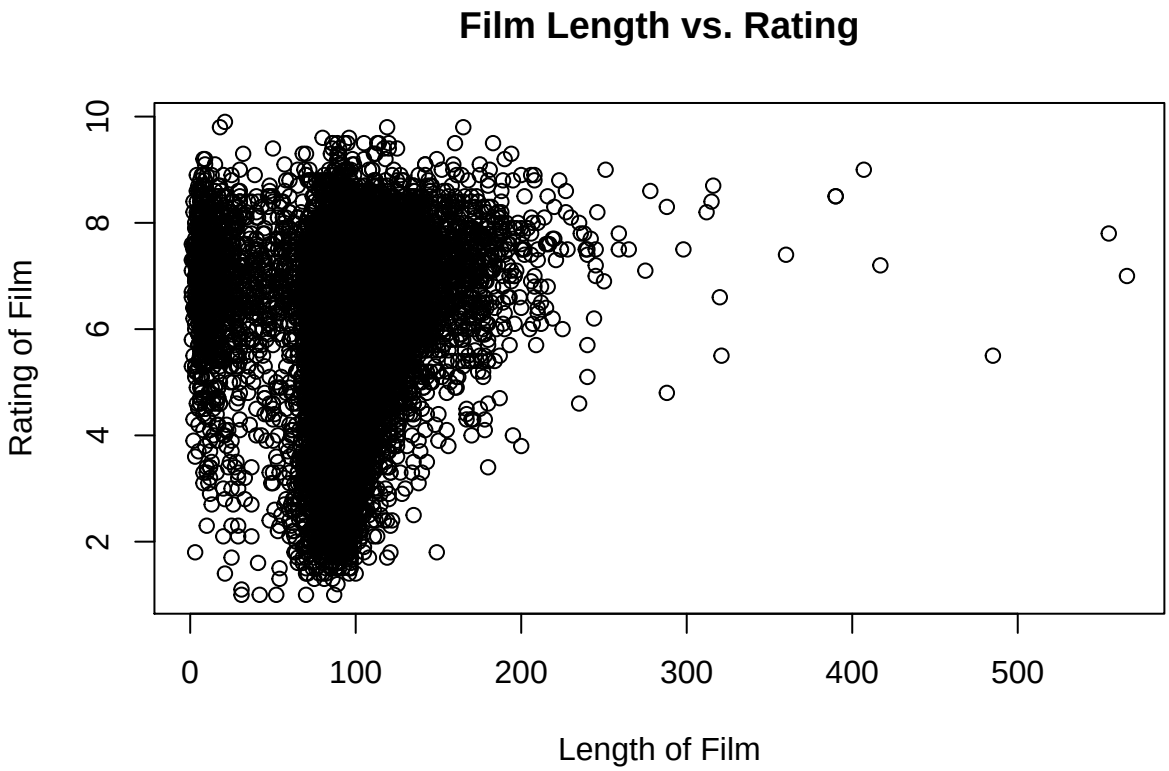
The dataset also contains an outlier for the variable 'length'. Below, I will plot a scatterplot of length and reception.

Figure 2



After observing the scatterplot showing the relationship between reception and length, it is clear that we have an outlier (a film titled 'A Cure for Insomnia' that is over 5000 minutes long — fitting title). Below, I will display what the data looks like without the outlier.

Figure 3



Removing the outlier pertaining to the film 'A Cure for Insomnia' ensures that we can run a good analysis without skewed results due to an outlier.

To understand what the dataset looks like, I will display the first few rows of our dataset:

Table 1

```
## # A tibble: 6 x 6
##   title                    year length budget rating votes
##   <chr>                   <int>  <int>  <int>  <dbl> <int>
## 1 $                        1971    121     NA    6.4   348
## 2 $pent                    2000     91     NA    4.3    45
## 3 $windle                  2002     93     NA    5.3   200
## 4 '94 du bi dao zhi qing   1994     96     NA    5.9    53
## 5 'A' gai waak             1983    106     NA    7.1  1259
## 6 'A' gai waak juk jaap    1987    101     NA    7.2   614
```

The population in this case is every person that watched the film and can rate the film on a scale from 1 to 10. The population frame in this case is all people who have an IMDB account, watched the film, and can rate the film on a scale from 1 to 10. Entries in the dataset were found by searching through IMDB and collecting data on films if they had a known length and had been rated by at least one IMDB user.

The key features of the data is the breadth of films that the data offers. We get information about films from various eras of cinema ensuring that our data is a good representation of films in general. The advantage of the dataset arises from the fact that our data is well representative of film in general; this means that when we perform an analysis on the data, we know that our results are meaningful. Despite the fact that it is well representative of film, it's largely representative of Western cinema and missing out other large industries around the world; this can be viewed as a disadvantage because our data does not represent the entire film industry. If our data included films from the Eastern world, the results from our analysis may be different.

There are also potential biases that arise from the dataset. Due to the fact that our data is collected from the website IMDB, the results of our data are primarily collected from the 21st century, meaning many of the reviews are retrospective. The retrospective review of older films may not be reviewed the same way as a modern film from the 21st century as the reviewer may look evaluate the films from a different criterion.

Model

In this paper, I will be using linear regression to determine how budget, length and year of release can affect a film's reception.

A linear regression model can be explained by the equation

$$y = a * x_0 + b$$

In this case, y is the film's dependant variable, while x_0 is the independant variable; a is interpreted as the slope coefficient, while b is the y-intercept .

I want to determine whether the independant variable affects the dependant variable. The purpose of linear regression is to approximate the relationship between the dependant and independant variable with a straight line. The straight line is fitted to reduce error between all of the points on the scatterplot between the dependant variable ('rating') and the independant variable (differs) . After approximating the relationship with a straight line, we can use the slope co-efficient to determine whether there is a positive or negative relationship. A positive slope coefficent implies a positive relationship, while a negative slope coefficent implies a negative relationship. But in order to determine whether the relationship is actually statistically significant, we need to look at the p-value.

The purpose of a p-value is to determine whether the relationship is actually significant or if it just occured by chance. The p-value assumes that the null hypothesis — there is no relation between the independant and dependant variable — is true, and finds the probability of obtaining results that are more extreme than the null hypothesis in the dataset. If the p-value is large, then that indicates that the data supports the null hypothesis; if the p-value is small, then that indicates that the data does not support the null hypothesis. I will be using a significance level of 0.05 meaning that if the p-value is smaller than 0.05, I can conclude that there is a significant relationship.

Using the lm() function in R, I will create a linear regression model, and then I will look at the summary of the model to determine if my results are meaningful. The summary of the linear regression model provides both the slope coefficient which allows me to determine whether we have a positive or negative relationship, and it also provides the p-value for the slope coefficient so I can identify whether there is a significant relationship. If the p-value is less than 0.05, then I know that the independant variable affects has a significant relationship with the film's reception.

The dependant variable in this linear regression model will be 'rating' because I want to determine how the rating of a film is dependant on other independant variable such as budget, length, and year of release. Since I am analyzing a 'film's reception' in this paper, the variable that best encompasses reception is the average rating that IMDB users provide for the film, which is why I am using 'rating' as the dependant variable.

The independant variables I will be analyzing are:

- 'length'- this provides the length of the film in minutes and is simply provided by the variable 'length'
- 'year'- this is simply provided by the variable 'year' and indicates the year of release for the film
- 'budget'- this is simply provided by the variable 'budget' and indicates the budget of the film in US dollars

All 3 variables that I am analyzing are continuous, numberical variables.

Because I am identifying how a film's reception is affected by various independant variables, it makes sense to run linear regression as I want to see the cause and effect relationship between an independant and dependant variable. The relationships that I am analyzing are strictly between 2 variables — rating (dependant) and length, year or budget (independant).

Results

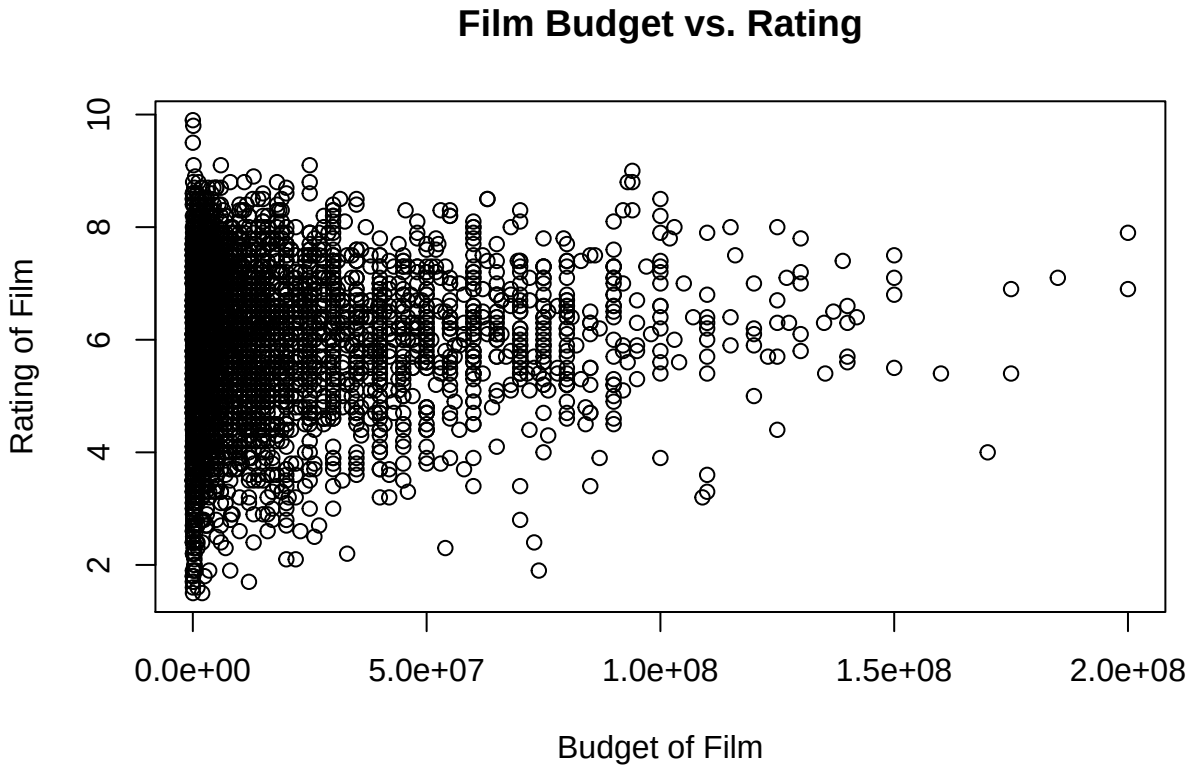
In this section of this paper, I will be looking at the relation between a film's rating and 3 different independant variables— length, year and budget. I will be using a signicance level of 0.05 to determine whether a p-value is significant.

How budget affects a film's reception

I will first look at how a film's budget affects the reception of a film

In order to determine the relationship between the dependant variable (reception) and the independant variable (budget), I will run a linear regression. Because many of the entries in our dataset don't contain information about the film's budget, I will remove those values when performing the linear regression; this means that the linear regression between budget and reception will only inspect entries with a budget greater than 0.

Figure 4



Looking at the data as a scatterplot, there seems to be slight positive relationship between the 2 variables.

To further examine the relationship, I will run a linear regression between the two variables to determine whether the relationship is significant. The results of the linear regression can be seen below:

Figure 5

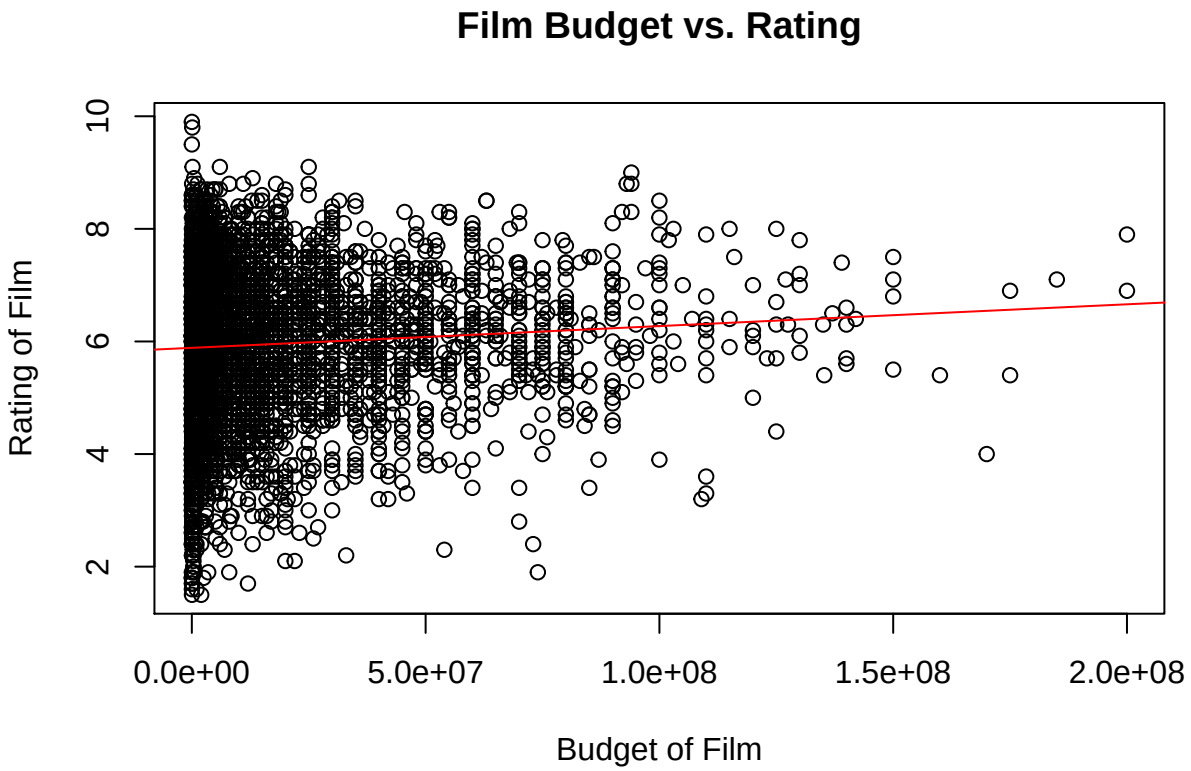


Table 2

```
##
## Call:
## lm(formula = rating ~ budget, data = budgetfiltered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3946 -0.8965  0.1358  1.0242  4.0131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.887e+00  2.721e-02 216.315  < 2e-16 ***
## budget      3.866e-09  8.654e-10   4.468 8.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.375 on 3815 degrees of freedom
## Multiple R-squared:  0.005205,   Adjusted R-squared:  0.004944
## F-statistic: 19.96 on 1 and 3815 DF,  p-value: 8.133e-06
```

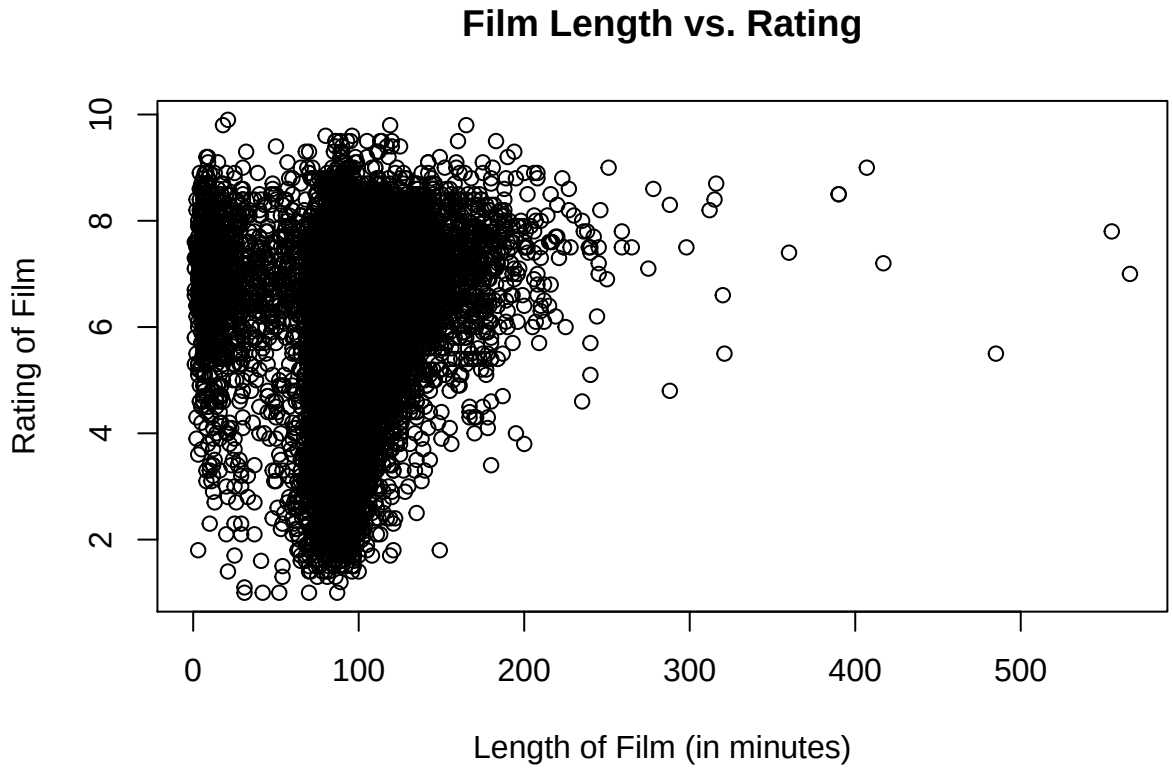
Judging by the linear regression summary, we can conclude that there is a meaningful relationship between the dependant variable, rating, and the independant variable, budget. This is due to the fact that the p-value (2e-16) is significantly smaller than the significance level of 0.05 we are using in this report.

How length affects a films reception

Next, I will be looking at how a film's length affects the reception of a film.

In order to determine the relationship between the dependant variable (reception) and the independant variable (length), I will run a linear regression, but first I will visualize the data with a scatterplot.

Figure 6



Although there is a large cluster on the left side of the graph, there does seem to be a slight positive relationship that indicates a longer length results in a better film rating. To confirm this, we can look at the summary statistics for the linear regression.

Figure 7

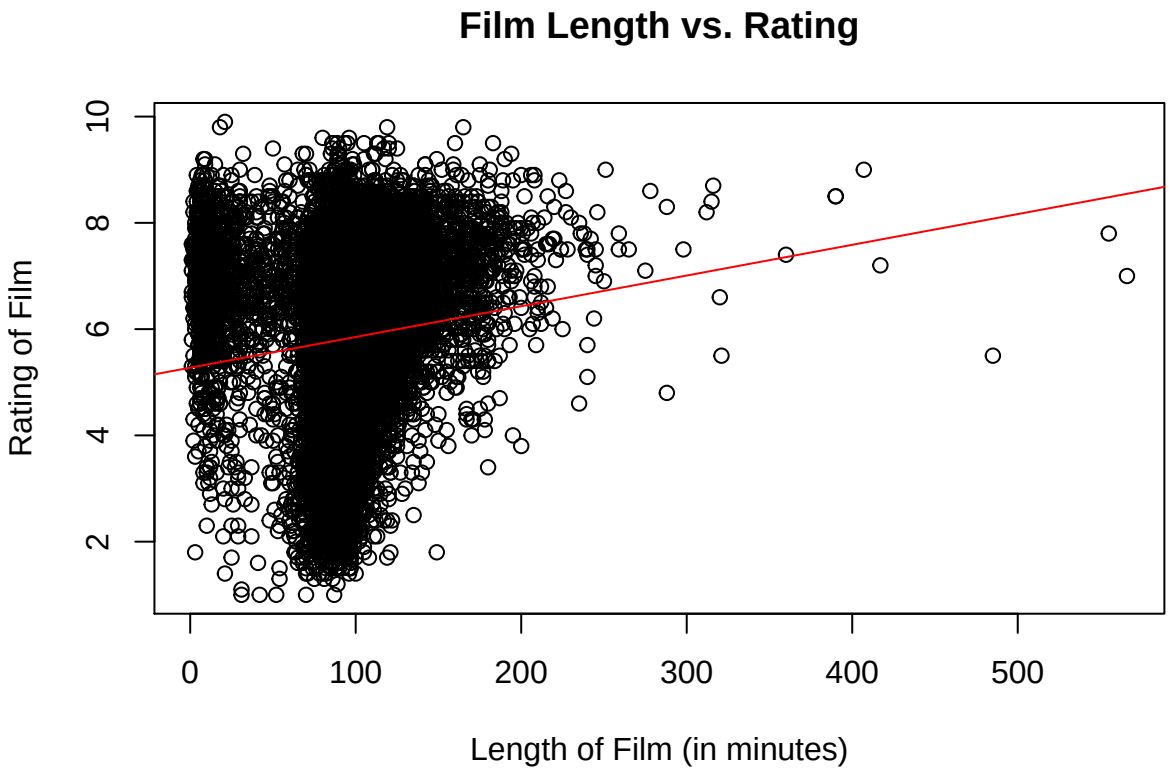


Table 3

```
##
## Call:
## lm(formula = rating ~ length, data = movieswithoutcureforinsomnia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7773 -0.9062  0.1534  1.0360  4.5043
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.2743050  0.0328375   160.6   <2e-16 ***
## length      0.0057812  0.0003285    17.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 24579 degrees of freedom
## Multiple R-squared:  0.01244,    Adjusted R-squared:  0.0124
## F-statistic: 309.6 on 1 and 24579 DF,  p-value: < 2.2e-16
```

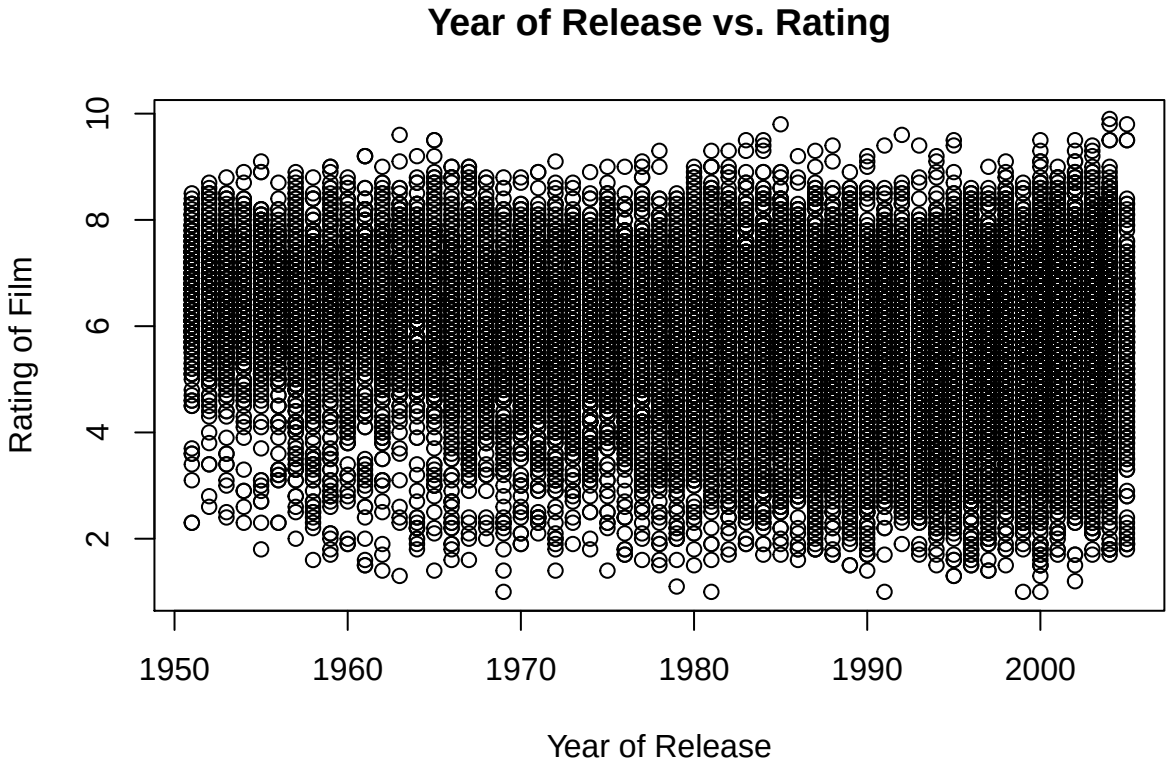
After observing the summary statistics related to the relationship between rating and length, it is clear that we have a really strong relationship. The p-value of 2e-16 is significantly smaller than our significance level of 0.05 indicating that there is a strong relationship between the variables rating and length.

How the year of release affects a films reception

Last, we will look at how the year a film's released affects it's reception.

To begin, we will visualize the data with a scatterplot to identify whether we can see a relationship.

Figure 8



Judging by the scatterplot, there doesn't seem to be an obvious relationship between the two variables.

We can accurately identify whether there is a relationship by examining the linear regression between the two variables.

Figure 9

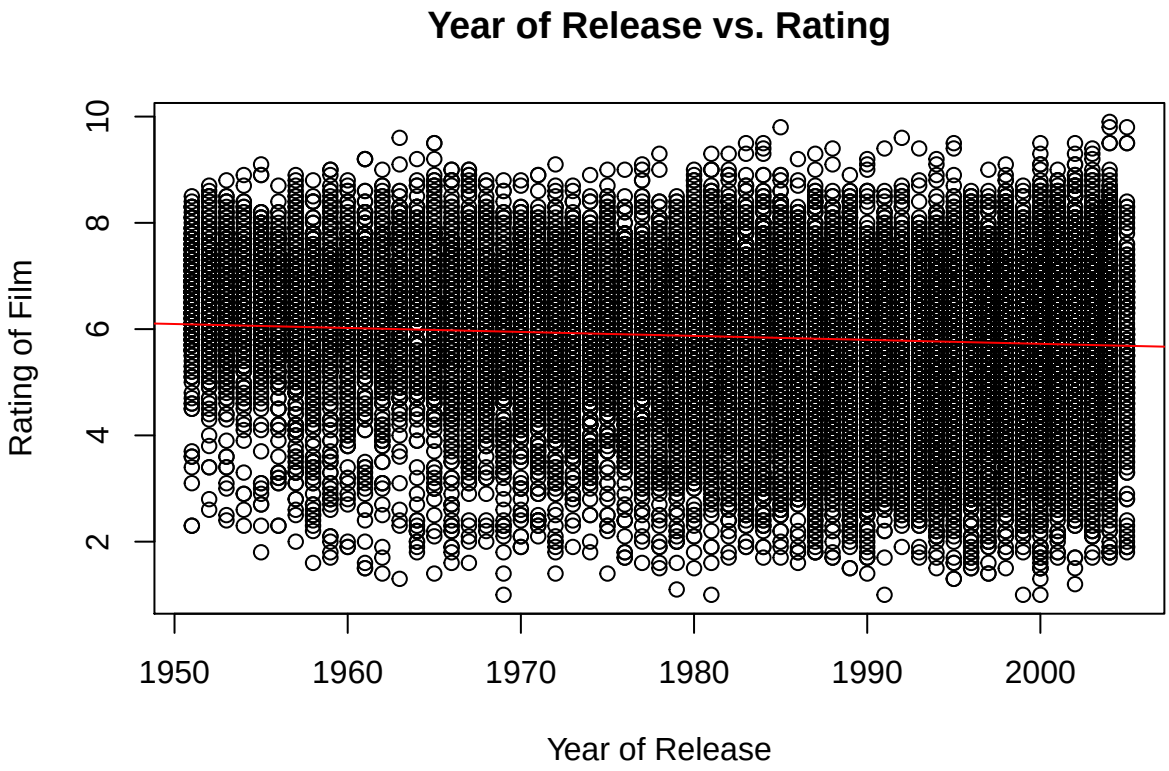


Table 4

```
##
## Call:
## lm(formula = rating ~ year, data = moviesfilteredx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9546 -0.9084  0.1618  1.0543  4.2065
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.6486986  1.1720001   17.62   <2e-16 ***
## year        -0.0074627  0.0005902  -12.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.419 on 24580 degrees of freedom
## Multiple R-squared:  0.006463,   Adjusted R-squared:  0.006422
## F-statistic: 159.9 on 1 and 24580 DF,  p-value: < 2.2e-16
```

Surprisingly, there is a strong negative relationship. By observing the graph, I thought that there wasn't a significant relationship as I didn't see any blatant positive or negative relationship. But the slightly negative slope with a p-value much smaller than 0.05 indicates that there is a negative relationship between year and reception.

Discussion

From this results of the analysis, I can conclude that there is a positive relationship between reception and budget, a positive relationshop between reception and length, and a negative relationship between reception and year of release.

The positive relationship between a film's reception and budget indicates that as a film's budget increases, the reception of the film increases as well. Many of the film's released before 1950 had positive reception, despite the low budget; this can be a result of the fact that many IMDB reviewers rated these films based on a set of criteria that they may have not used for films released after 1950. This criterion may place a heavy emphasis on what filmmakers were able to do with technology and norms relative to that time. Because of this dichotomy between the seperate eras, I decided to focus exclusively on the era after the 1950s and discovered that there is a positive relationship between a film's reception and budget.

A larger budget means that the film production has more money to spend; this can results in better actors, special effects, marketing, etc. which can affect how an audience reacts to the film. Despite the fact that a film production may have more money to spend, it doesn't always result in an overall better film. Critics may interpret a good film differently than what an audience interprets as a good film.

There are also some issues that arise from simply quantifying budget in US dollars. Inflation plays a large part in why films in the 1990s to 2000s have larger budgets. If earlier films like those from the 1950s were adjust for US inflation, then our results may have different and we may have discovered that there actually is no relation.

The positive relationshop between a film's reception and length indicates that as a film's length increases, then the reception of the film increases as well. Films released before 1950 were often characterized by their short lengths (especially during film eras before the 1920s). Due to their short lengths and the fact that cinema wasn't used in it's typical way at that time, I decided to focus excluvely on films released after the 1950s.

Film studios often have a final say before releasing a film. If a film is long, that may advise cutting the film so that it appeals to the masses. For a film to have a long runtime, it would need to be justified; the film would have to be good enough to be worth the audience's time. This may be a reason why a longer run time leads to a better reception. Good filmmakers often want their film to be a result of their uncompromised —without cuts— vision leading to more films that are longer with a better reception.

The negative relationship between a film's reception and year indicates that older films are likely to have better reception than newer films. This relationship could be a result of a multitude of factors. As we approach the 2000s, creating a film did not have many barriers as it did many years ago. During the 1950s, to create a film was a lot more expensive due to expensive, heavy, equipment, whereas in the 2000s, film could be be made on a lightweight, cheap video camera.Because of the plight many filmmarkers faced during the 1950s, it may have resulted in a more well-thought out —and therefore, better— films.

The data was collected from IMDB meaning that a lot of earlier films were rated retroactively. This may result in an inflated score for some of the older films. Many people adopt an 'older is better' mentality when watching older films, which causes them to rate older films with a higher score compared to newer films.

The results from the analysis can assist people in many ways. When deciding on a good film to watch, the positive relation reception has with budget and length can help identify whether a film may be a good or a bad film; the negative relation between reception and year of release can also aid in helping determine the film's reception. Additionally, when creating a film, some of these relationships can aid in forecasting it's reception. A longer film, with a large budget may result in a film that is well received.

Advantage

Disadvantage IMDB is a newer site so we are not seeing reviews from people now.

The results of our analysis do contain weaknesses, many of them stem from the data being used.

There are many additional steps that can be taken for a better analysis. For one, we could have a dataset that contains information on more films from around the world. Have budget be adjustted for inflaiton.

Appendix/References

Code

Below you will also find the R code used in this analysis:

Preamble; Setting up libraries and filtering the dataset

```
# Setting up libraries
library(ggplot2movies)
library(ggplot2)
library(dplyr)


#Filtering the dataset so that it only contains relevant variables
moviesx <-movies[c('title', 'year', 'length', 'budget','rating','votes') ]
#Filtering the dataset so that it only contains films that contain a certain amount
#of votes. Specifically, above the 50th percentile which is 30
moviesfiltered <- filter(moviesx, votes > (quantile(movies$votes, .5)))
#Filtering the dataset so that it only contains films released after the year 1950
moviesfilteredx <- filter(moviesfiltered, year > 1950)
```

Figure 1

```
#creating histogram of films based on year they were released
hist(movies$year, main = "Histogram of Film Years", xlab= "Year of Film Release")
```

Figure 2

```
#Showing what the scatterplot between rating and length looks like without removing
#outlier
plot(rating~length,data=moviesfilteredx , main = "Film Length vs. Rating", xlab =
       "Length of Film", ylab = "Rating of Film")
```

Figure 3

```
#filtering dataset so that it does not include outlier
movieswithoutcureforinsomnia <- filter(moviesfilteredx, length < 3000)
#Showing what the scatterplot betweem rating and length looks like without the outlier
plot(rating~length,data=movieswithoutcureforinsomnia , main = "Film Length vs. Rating",
     xlab = "Length of Film", ylab = "Rating of Film")
```

Table 1

```
#Code to show first few rows of data
head(moviesfilteredx)
```

Figure 4

```r
#Remove entries that have a value of NA or 0 for budget
budgetfiltered <- filter(moviesfilteredx, budget > 0)
#Scatterplot of rating and budget
plot(rating~budget,data=budgetfiltered  , main = "Film Budget vs. Rating",
     xlab = "Budget of Film", ylab = "Rating of Film")
```

Figure 5

```r
#Creating a variable for the linear regression
linregbudget <- lm(rating~budget, data=budgetfiltered)

#plotting graph with visualization of linear regression
plot(rating~budget,data=budgetfiltered , main = "Film Budget vs. Rating",
     xlab = "Budget of Film", ylab = "Rating of Film")
abline(linregbudget, col="red")
```

Table 2

```r
#Summary statistics for linear regression
summary(linregbudget)
```

Figure 6

```r
#Scatterplot of rating and length
plot(rating~length,data=movieswithoutcureforinsomnia,
     main = "Film Length vs. Rating", xlab = "Length of Film (in minutes)",
     ylab = "Rating of Film")
```

Figure 7

```r
#Variable for linear regression between rating and length
linreglength <- lm(rating~length, data=movieswithoutcureforinsomnia)

#Visualizing scatterplot with linear regression overlayed
plot(rating~length,data=movieswithoutcureforinsomnia,
     main = "Film Length vs. Rating", xlab = "Length of Film (in minutes)",
     ylab = "Rating of Film")
abline(linreglength, col="red")
```

Table 3

```r
#Summary statistics for linear regression
summary(linreglength)
```

Figure 8

```r
#Scatterplot of rating and year
plot(rating~year,data=moviesfilteredx, main = "Year of Release vs. Rating",
     xlab = "Year of Release", ylab = "Rating of Film")
```

Figure 9

```
#Variable for linear regression between length and year
linregyear <- lm(rating~year, data=moviesfilteredx)

#Visualizing scatterplot with linear regression overlayed
plot(rating~year,data=moviesfilteredx , main = "Year of Release vs. Rating",
     xlab = "Year of Release", ylab = "Rating of Film")
abline(linregyear, col="red")
```

Table 4

```
#Summary statistics for linear regression
summary(linregyear)
```

References

- Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. [ggplot2]
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr [dplyr]
- Hadley Wickham (2015). ggplot2movies: Movies Data. R package version 0.0.1. [ggplot2movies] https://CRAN.R-project.org/package=ggplot2movies
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. [R]