How film length, year and budget affect a film's reception among audiences

Arjun Dhatt

20/12/2020

Keywords

- Film
- Genre
- Budget
- Linear Regression
- Title

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 varaibles such as — title, rating, genre, and budget. Then I will run an analysis to identify the variables that affect a film's reception among audiences. After completing the analysis, I concluded that a film's length and year of release do affect a film's reception, while the budget does not.

Introduction

Many factors can influence an audience's reaction to a film. Various variables such as the length of the film, the genre, budget can influence the film's reception. This poses the interesting question — how does a films length, genre and budget affect it's reception among audiences?

In the report, I will investiage 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 — film title, film length, film rating, anf film budget.

Since some of the variables I listed above do not contain information for some important 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 variable budget. I conducted a linear regression and using p-value I determined that there is no relationship. Next, I looked at variables length and year and surprisingly discovered that there is a relationship.

With the results from the analysis discussed above, I concluded that a film's length and year of release do affect the film's reception, while the budget does not.

In the rest of this report, I will begin by looking at the data, showing my analysis, and then a conclusion.

Data

```
library(ggplot2movies)
library(ggplot2)
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
```

```
movies <- mutate(movies, decade = ifelse(year %in% 1950:1959, "1950s",
                                         ifelse(year %in% 1960:1969, "1960s",
                                         ifelse(year %in% 1970:1979, "1970s",
                                         ifelse(year %in% 1980:1989, "1980s",
                                         ifelse(year %in% 1990:1999, "1990s",
                                         ifelse(year %in% 2000:2009, "2000s", "<1949")))))))

movies <- mutate(movies, )

moviesx <-movies[c('title', 'year', 'length', 'budget','rating','votes') ]


moviesfiltered <- filter(moviesx, votes > (quantile(movies$votes, .50)))
```

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; not all films contain information about all 25 categories. For example, many of the films from before 1960 do not contain information about the film budget which will be 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
- 'mpaa' — provides the MPAA (Motion Pictures Association of America) rating of the film

To make the variable 'year' easier to analyze, I created a variable titled 'decade' so that the films are grouped by decade. This makes it much easier to analyze because there are fewer values to analyze.

Below, you can see what the first few values of our dataset looklike:

```
head(moviesfiltered)
```

```
## # 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
```

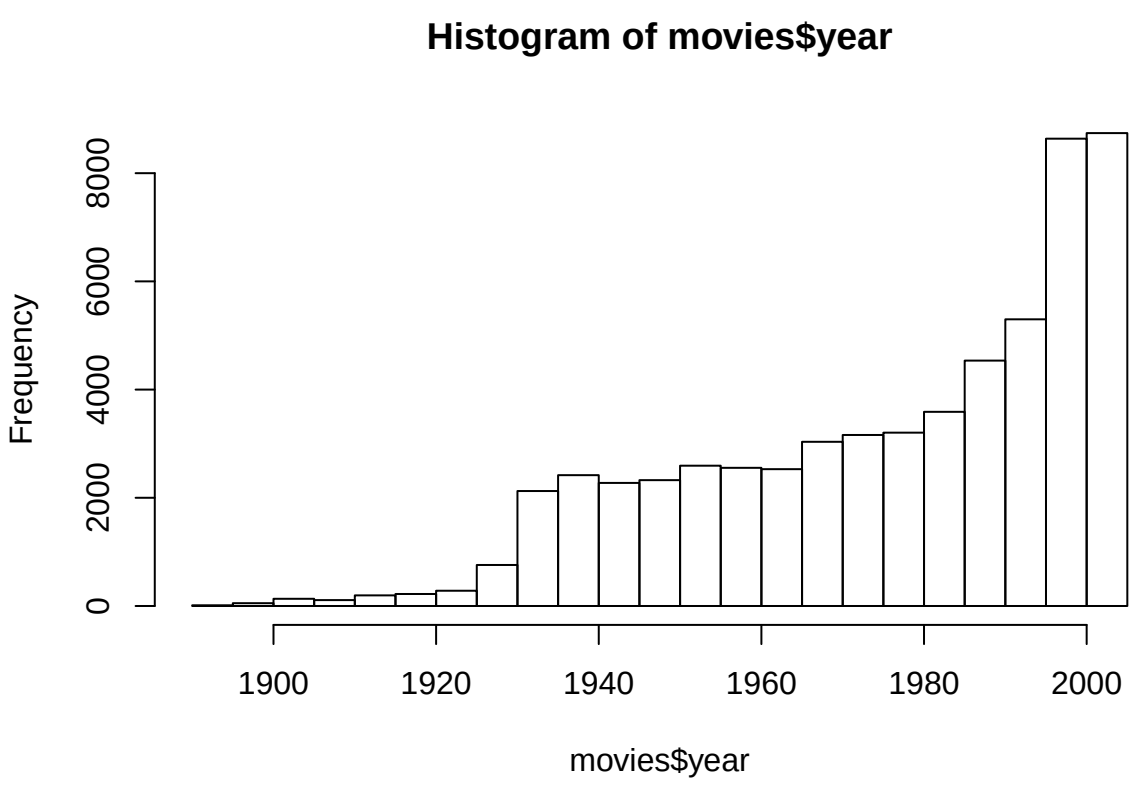
```
## 3 $windle                  2002     93     NA    5.3   200
## 4 '49-'17                  1917     61     NA    6      51
## 5 '94 du bi dao zhi qing   1994     96     NA    5.9    53
## 6 '?' Motorist, The        1906     10     NA    7      44
```

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.

When observing the dataset, we can see that the dataset is severely left-skewed by looking at a histogram based on a film's year of release.

```
library(ggplot2movies)
library(dplyr)
hist(movies$year)
```



Model

In this paper, I will be using linear regression to see how different variables can affect a films reception.

A linear regression model can be explained by

$$y = a * x_0 + b$$

. In this case, y is the film's dependant variable and details the films reception, while x_0 is the independant variable that I want to determind if it affects the dependant variable. 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 see if the p-value is significant. If the p-value is less than 0.05, then I know that the independant variable affecst a films reception.

The dependant variable that I will be using in the linear regression model is 'rating'. 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.

The independant variables I will be analyzing are: - 'length'- this is simply provided by the variable 'length' - 'year'- this is simply provided by the variable 'year'. I did mutate another variable titled 'decade' which makes visualizing the data easier, but when performing linear regression, 'year' is more thorough so I decided to use 'year as an independant variable. - 'budget'- this is simply provided by the variable 'budget'

All 3 variables that I am analyzing are continuous 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.

Results

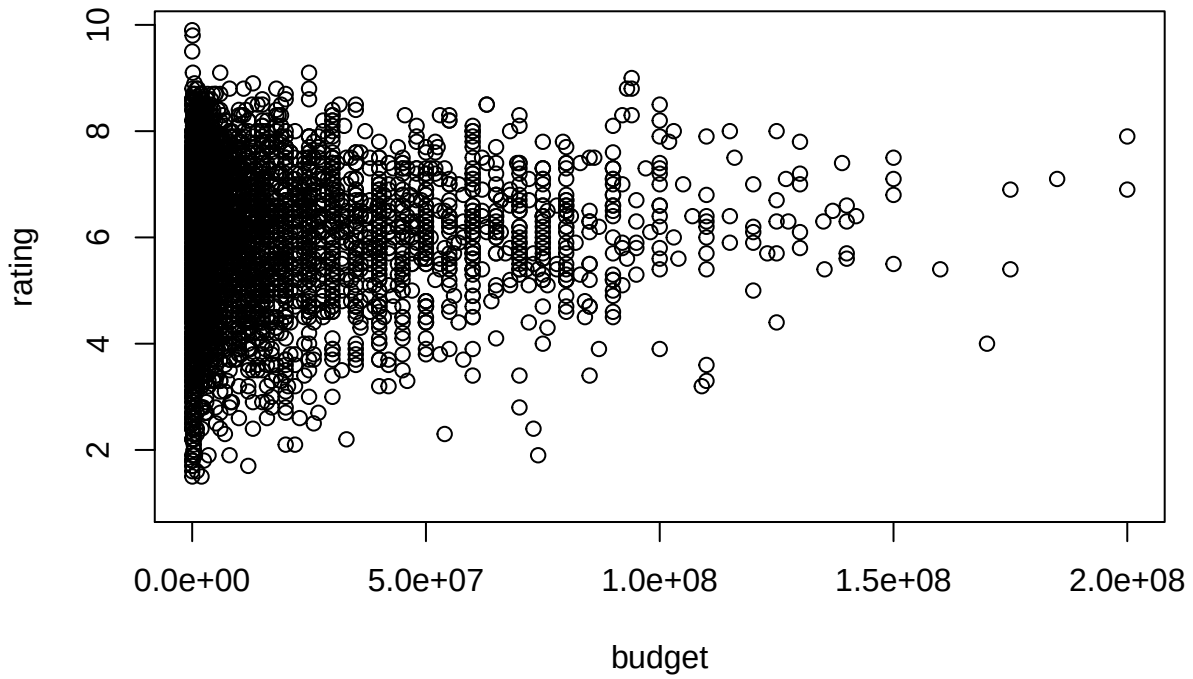
How budget affects a films reception

We will first determine how a film's budget affects the reception of a film. By running a linear regression on the variables, we can see what a scatterplot between the variables looks like:

```
#Many rows of our data do not contain a budget, so I will remove those rows
budgetonly <- na.omit(moviesfiltered)
linregbudget <- lm(rating~budget, data=budgetonly)
summary(linregbudget)
```

```
##
## Call:
## lm(formula = rating ~ budget, data = budgetonly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5420 -0.8426  0.1588  1.0456  3.8604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.040e+00  2.496e-02 241.986   <2e-16 ***
## budget      1.196e-09  8.395e-10   1.425    0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.365 on 4270 degrees of freedom
## Multiple R-squared:  0.0004752,  Adjusted R-squared:  0.0002411
## F-statistic:  2.03 on 1 and 4270 DF,  p-value: 0.1543
```

```r
plot(rating~budget,data=moviesfiltered )
```




```r
#Judging by the high p-value, we can conclude that a films budget does not affect its reception among t
```

Judging by the scatterplot, we can see that there is not a clear relation between the variables. By further analyzing the linear regression summary, we can see that we have a large p-value and can conclude that there is no relation between a film's reception and budget.

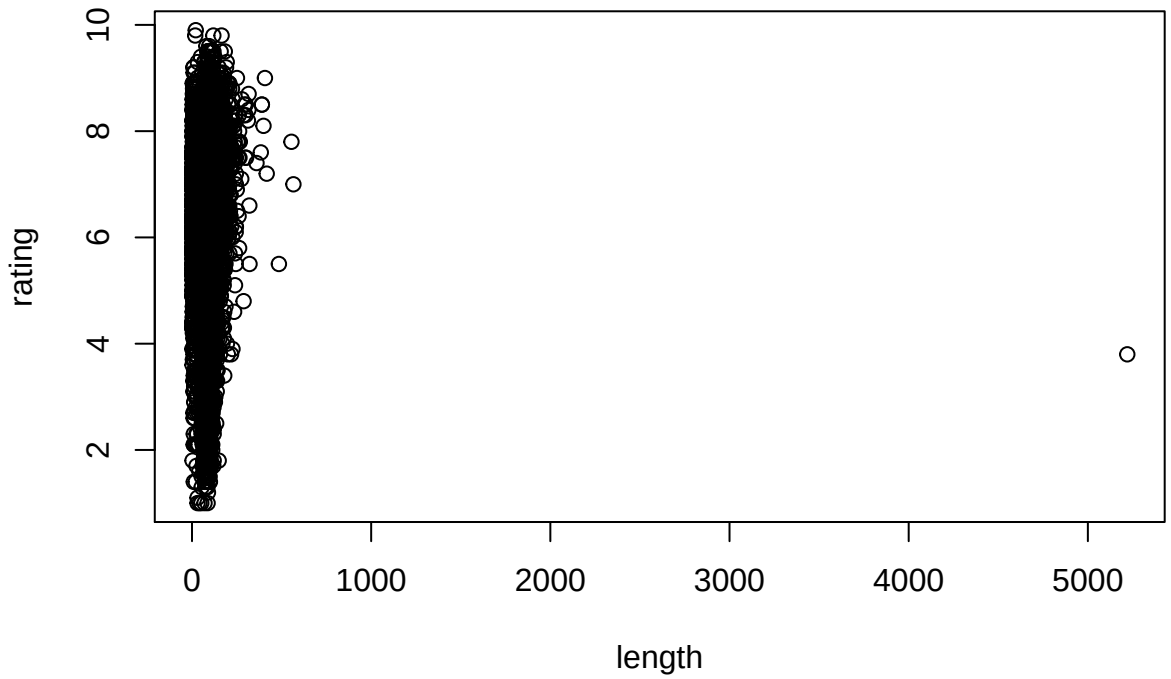
How length affects a films reception

Next, we will determine how a film's length affects the reception of a film. By running a linear regression on the variables, we can see what a scatterplot between the variables looks like:

```r
linreglength <- lm(rating~length, data=moviesfiltered)
summary(linreglength)
```

```
##
## Call:
## lm(formula = rating ~ length, data = moviesfiltered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9570 -0.8594  0.2275  1.0353  3.9750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.9148270  0.0195429 302.658   <2e-16 ***
## length      0.0004844  0.0001910   2.536   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 28934 degrees of freedom
## Multiple R-squared:  0.0002222,  Adjusted R-squared:  0.0001876
## F-statistic: 6.431 on 1 and 28934 DF,  p-value: 0.01122
```

```r
plot(rating~length,data=moviesfiltered )
```




```r
#Yes, films that are longer get better rating
```

Judging bu the scatterplot, we do see a slight relation between a films length; it seems as if though the longer the length of the film, the higher the reception. Since our p-value is <0.05, we can conclude that there is a relation between the 2 variables.

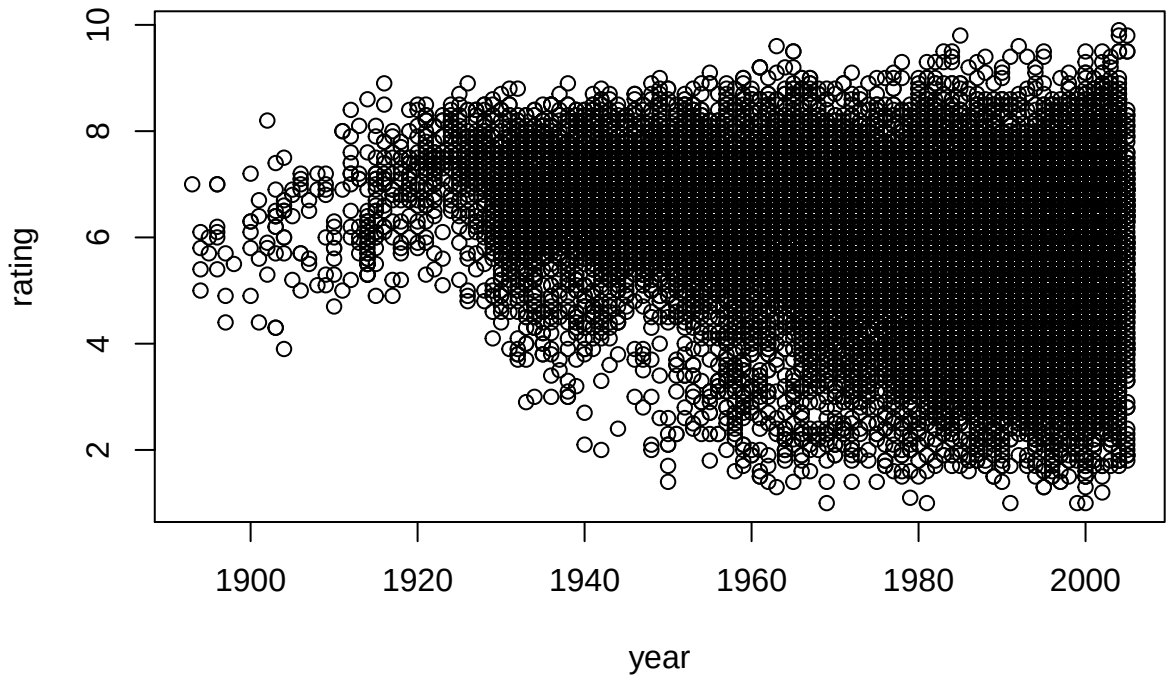
How the year of release affects a films reception

Last, we will determine how the year a film is released affects the reception of a film. By running a linear regression on the variables, we can see what a scatterplot between the variables looks like:

```r
linregyear <- lm(rating~year, data=moviesfiltered)
summary(linregyear)
```

```
##
## Call:
## lm(formula = rating ~ year, data = moviesfiltered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0880 -0.8273  0.1468  0.9800  4.2807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.4598534  0.7081523   45.84   <2e-16 ***
## year        -0.0133935  0.0003579  -37.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.365 on 28934 degrees of freedom
## Multiple R-squared:  0.04617,    Adjusted R-squared:  0.04614
## F-statistic:  1401 on 1 and 28934 DF,  p-value: < 2.2e-16
```

```r
plot(rating~year,data=moviesfiltered )
```




```r
#Yes, films that are longer get better rating
```

In the scatterplot, we can see that there is a clear relation. Obviously the yaer 2000 has a larger interval of rating and that may be partly due tot the fact that there was more movies released. Judging by the fact that the p-value is <0.05 in this case, we can conclude that there is a relationship among the year a film was released and its reception. ## Discussion

Show what we concluded from our analysis

Appendix/References

Show all lines of code and extranneous graphs