Python: Data Science

Christopher Darrall

QA Consulting

Python: Data Science

An Analysis of Film Ratings By Length



Contents

[Hypothesis 2](#_Toc518053177)

[Dataset 2](#_Toc518053178)

[Considerations and Issues 2](#_Toc518053179)

[Methodology 2](#_Toc518053180)

[Results 3](#_Toc518053181)

[High Budget Films 3](#_Toc518053182)

[Mid Budget Films 3](#_Toc518053183)

[Low Budget Films 4](#_Toc518053184)

[Results and Potential Explanations 4](#_Toc518053185)

# Hypothesis

The Hypothesis is as such:

Films that are longer will generally have a higher score

# Dataset

The dataset was sourced from the movies\_metadata.csv from the Kaggle dataset repository by [Rounak Banik](https://www.kaggle.com/rounakbanik/the-movies-dataset). The Dataset was pulled from IMDB’s database and contained 45467 movies.

## Considerations and Issues

A large number of the movies had incomplete data and as such below 4000 films where valid for the study after films were removed based upon:

* Having No Budget Recorded: As these films would not be usable from the offset.
* Having a budget below $2.5million: This was to filter out self-produced films (i.e. *The Room*) and potential cash-grabs (i.e. *Transmorphers*). This would also eliminate a number of shorts and concept films that would skew the data as they would be rated on exploiting a single element (i.e. *Loving Vincent* for its art and *Dear Basketball* for its sentimentality).
* Having Less than 100 rating votes: To prevent overly skewed films due to limited ratings. Such as films that where only seen/rated by friends and families and other Obscure films.

There are issues with my considerations, for instance they effectively eliminate the precence of many Cult Classics (i.e. *The Rocky Horror Picture Show,* budgeted at $1.4 million) and other highly rated films from alternative backgrounds (i.e. The Ugandan film *Who Killed Captain Alex?*, which has a rating of 8.2 at 62 minutes long is excluded due to its budget of $200). Another issue is that films are not adjusted according to inflation.

# Methodology

I broke the films up into three sections first so as to be able to better track the results and to prevent the trend line from being obscured. The three categories where Low Budget films, Mid Budget films and High Budget films. The intent is not to follow industry standard definitions but to break up the study to allow for better analysis.

# Results

## High Budget Films

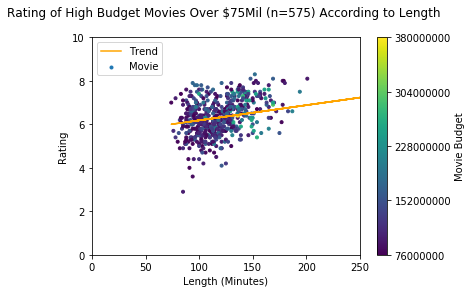


Figure 1: Rating of High budget movies according to length

High budget films has a slight trend of getting a higher rating based upon length, the majority of movies tended to cluster towards the lower end of the budget at $75 million, with the summer blockbusters like *Avengers: Age of Ultron* raising the upper limit to $380 million. Despite the large range of film budgets this is the smallest data set.

## Mid Budget Films

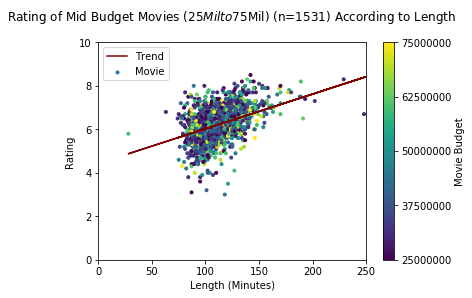


Figure 2: Rating of Mid budget movies according to length

Mid Budget films had a strong trend of getting a higher rating based upon length. The range for medium budget films being from $25million to $75million was to capture films that would have the budget to competently afford a full production of staff while avoiding films that carried of star power and summer blockbusters that may be biased by “*hype*”. There is a good distribution of budgets and there are more noticeable spread and outliers. Especially with regard to time (possible being caused by lower budget or older Epics, such as *Arn: The Knight Templar*).

## Low Budget Films

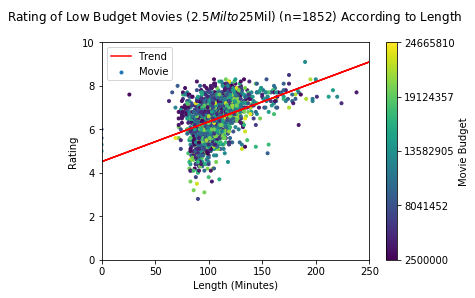


Figure 3: Rating of Low budget movies according to length

Low Budget films had the strongest trend of getting a higher rating based upon length, it has the most films of the three sections. There are even more outliers in this section, possibly because the lower budgets represented can produce high production shorts, while simultaneously mean that even average length films would have budgeting concerns.

## Results and Potential Explanations

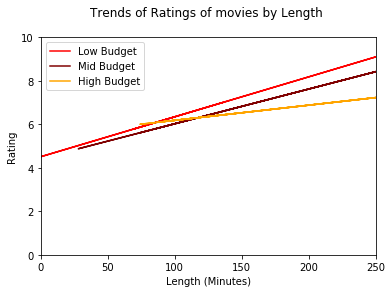


Figure 4: A comparison of the Trends by Budget

I believe that the results show that my hypothesis is true. However, there is not a single clear explanation as to why.

For example, with High budget films they overall have good scores however there a number of outliers. With films like *Gigli* that suck beyond the normal parameters of a film. This is often because if a high budget film is being pushed through and begins to fail companies will push more and more money into it to try and make it a success, and even if it is terrible they will want to try and recoup some money. High budget films are also lifted by Summer Blockbusters and spectacle masterpieces like *The Lord of the Rings trilogy* which tend to be longer than normal films. The overall lack of outliers would possibly be explained by large budget films often working to formula to minimize the odds of a flop because of the big risk with the budget.

As stated before There is a better representation of higher ratings in the Lower Budget films than the Mid Budget films, perhaps because as the critic Ben Crowshaw once said in regards to the success of cheaper Star Wars Originals against its mega budget prequels,

*“Limitations breed* creativity”

So while low budget films may be represented by a number of “bad” film makers it is also represented by up and coming directors and/or directors who have a limited budget developing a passion project that the studio has little faith in, a good example of all these factors is Makoto Shinkai’s *5 Centimeters Per Second*, which is 50 minutes long and was produced for $5 million, while having a favourable score of 7.7.

The lack of budget for the mid and low budget films also adds to the sharp incline as if a film is poor the studios will often “chop it up” to try and improve its chances, often working by cutting scenes they deem boring to create a faster paced film, while the higher rated films may of won over the support of their production company. If the budgets of these films where to run out then its game over, without an injection from studios a large number of films will be rushed to finish production, shortening the film and damaging its potential score, (such as *Highlander 2* (4.1 rating) *which* ran out of budget quickly due to the collapse of the Argentine economy that they were relying on having converted a large portion of the budget to pesos. This caused the studio to heavily edit the film *which was already awful*).

With all three sections, there is typically a pronounced cluster towards the lower ratings at the 90-minute mark may be due to the dearth of generic, movies that often pad to the arbitrary 90-minute mark for no particular reason. Often to be forgotten in the month after release, (I have already forgotten about 2018’s *Rampage,* and I watched it)