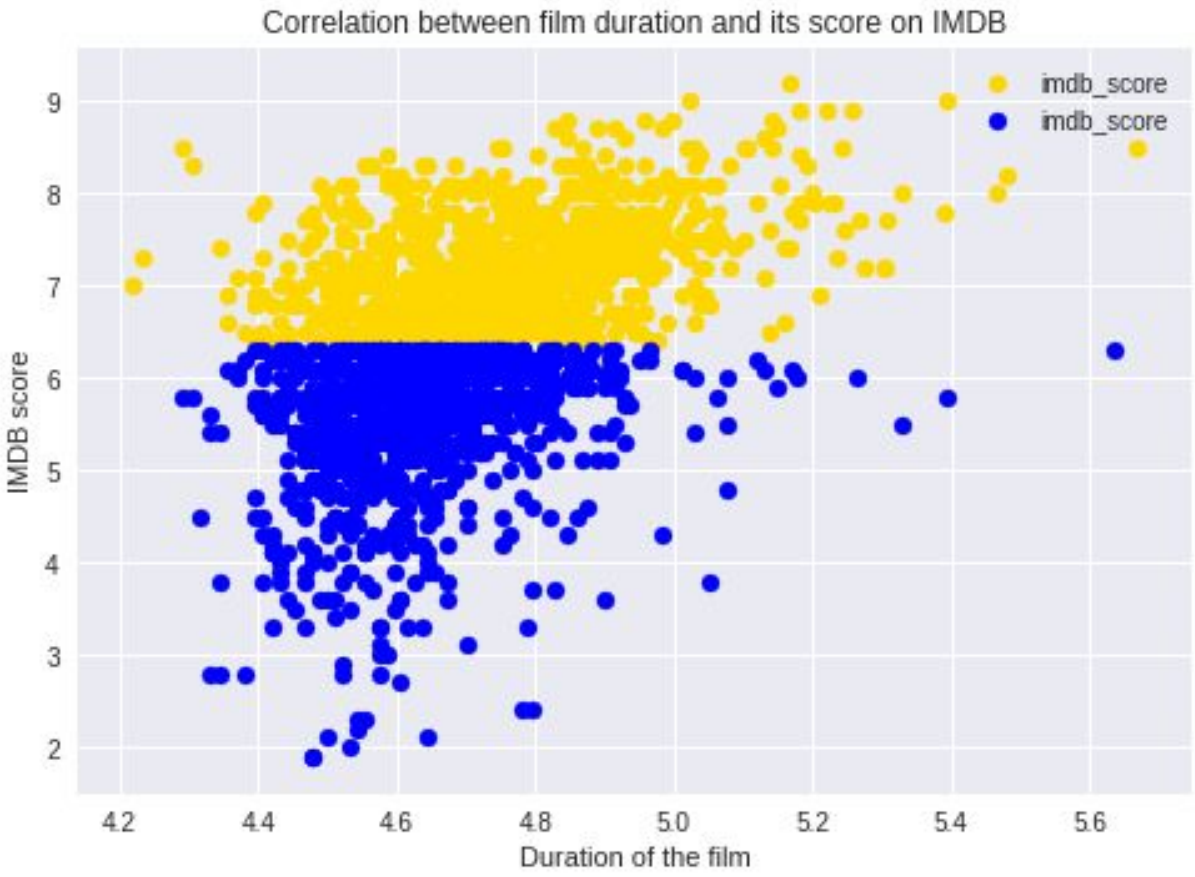
Analysis of successful vs unsuccessful films

In an attempt to discover what makes a film successful



Myroslava Romaniuk

FALL SEMESTER
DATA PREPROCESSING

DATA ANALYSIS

The idea for this project appeared unexpectedly. I like film, and yet a lot of films that I consider to be good do not make it into the popular category. So I decided to look at the film data and see what does a film need to be successful.

For the first part, I did a regular exploratory analysis of the IMDB 5000 movie dataset to better understand the data I was dealing with and to determine the points based on which I would separate successful from unsuccessful films.

For the second and main part, having determined which films are normally considered successful and which are not, I started analysing the dataset to see what role things like budget, film duration, film genre or film director play in the film's success.

[To lessen confusion from now on I will refer to the dataset of all the movies (after cleaning up duplicate rows, documentaries, irrelevant columns and making it numeric-only) as the original dataset.]

When analysing correlations between columns in the original dataset it seems that the film score is quite noticeably affected by duration and even title year, and very little by things like budget (what is worth noting is that the title year correlation was negative, meaning people liked older films better). Film gross was correlated with how many users/critics voted for it (hence also how many people saw it), and also duration and title year.

From the data it could also be seen that the fluctuation of scores has increased over the years (now there is a bigger gap between max and min scores) which could be partly explained to the ever-growing number of films and perhaps a certain lack of quality for quite a few of them (all illustrated in the notebook).

Now as for process deciding which movies were successful and which were not I decided to go with the approach the film industry also has. First off, to get a (roughly) even number of films on both sides, I decided that each film that has a score of 6.4 and above is well-rated, and everything below 6.4 is not. Out of the ones that are a 6.4 and higher I only took films that grossed at least $30m because

everything below that could well be considered niche, whereas I am interested in more or less mainstream films, and that additionally I decided that successful films made more money than was their budget. Unsuccessful were less than 6.4 and made less than the company spent making them. (Additionally, I only considered movies made in the last 50 years, in order to not really go into the "very old" film territory.)

RESULTS

I compared the budget of successful vs unsuccessful films to see whether how much it cost to make a film actually had any influence on the success of the movie. And no, it did not. Furthermore, in the last 20 years unsuccessful films have turned out to be more expensive. Sure, part of that is because my initial criteria for an unsuccessful film was that the studio spent more than it earned, but also since all of the unsuccessful films are first and foremost poorly rated, the outtake of this is that the studios throw away money making poor quality productions that do not even appeal to viewers.

The next part (exploring connections between duration and score, gross and score, and duration and gross) revealed the following: there is not much correlation between duration and IMDB score (although yes, longer movies appear to be more highly rated but also there are much less of them as a whole), but highly rated (>=6.4 IMDB score) films are shown to make a good amount of money, whereas the results among poorly rated films vary greatly. One more outtake is that films with below-average and average duration tend to make more money than the especially long ones.

As for genres, adventure, animation, biography, drama, history, family and sport films tend to be more successful than action, comedy and horror films (and the rest are not much different), even though there is still nothing critically different. At best, this could just serve as a warning to more carefully approach certain genres while at the same time trying to create content of better quality.

Analysing director data, one can tell that some directors are more likely to create a successful movie than the others, although many directors who have quite a few films might not actually have any successful ones. Also apparently the biggest

number of successful films was made a few years ago, and the biggest number of unsuccessful films was made in the early 2000s.

Having gotten so much information from the movie dataset, there is still not one clear thing that influences film success. It seems it must be a combination of a large amount of factors, at least one of which has to be pure luck. It was cumbersome to find a dataset that would provide all the information needed to further research. Ideally, for figuring out what might have a role in a film's success, I would be interested to see on what (exact) dates those movies came out, what political events might have surrounded them, who were the studios that made them, how much the production spent on music or special effects out of the whole budget, etc. As it is, this research still manages to provide an answer (albeit perhaps an unsatisfying one). What makes a film successful? Apparently, we just don't know. And neither do the studios making those films.

REFERENCES

1. Hitchhiker's guide to Exploratory Data Analysis: https://towardsdatascience.com/hitchhikers-guide-to-exploratory-data-analysis-6e8d896d3f7e
2. Matplotlib documentation: https://matplotlib.org/index.html