Aaron Drexler

DSC 680

Term Project Proposal 1

Topic: Movie correlations

I want to know the correlation between the top three measures of a movie: Budget, Revenue, and Popularity.

Business Problem

_Often, you will hear that this movie underperformed or overperformed their statistics. But that concept is so vague, I wanted to try to look to clarify and expand on how key aspects and performance stats compare with one another and which movie would be considered a failure, just through the numbers.

Data Sets

For my data set, I found a list of approximately 10,000 different movies from a Kaggle source, listed below. This database contains popularity, budget, revenue, title, cast, website, director, tagline, keywords, overview, runtime, genre, production companies, release date, vote count, vote average, release year, and adjusted budget/revenue totals. I chose to focus on the former three topics, popularity, budget, and revenue, as they are most often used in reference to measure a movie success.

Methods

To be specific, there are three different things that I will be measuring. First, I will be looking for the correlations between the budget and the revenue of the film. We often find that different movies either outperform or underperform their budget, but I want to see if there is truly a correlation between these two numbers is or if they are more independent than most think. The next topic under consideration is the correlation between the budget of the film and its popularity. In other words, is it true that high budget films lead to more popular films? Lastly, I wanted to create a logistic regression to create a true idea of whether a movie is successful. I plan to do this using the profit of a movie (Revenue – Budget) As compared to its popularity.

Ethical Considerations

There are a few different considerations when looking at this project. First off, this is a purely numerical, populist ranking of the movies. It does not take into consideration the actual cinematic topics, but rather focused solely on the numerical performance of the films. Also, my numbers that I plan to use reward populist firms. While these films tend to have the highest profit and popularity, they may pander to audiences as compared to other smaller films, which have more meaning and depth, but are less appealing to all. Next, I am using the iMDB rating, which is just one measure of popularity. It does not factor in other movie ranking sites, such as rotten tomatoes, which may have different measures of popularity. Finally, I am reliant on third hand figures of budget and revenue, which may not be the most accurate data, for all I know.

Challenges/Issues

I don't foresee too many issues. One issue that may pop up is the use of profit in the logistic regression. It is a different calculation that I will need to add on and will result in negative totals. I worry that these negative numbers will affect the logistic function. Another issue is that the database is missing some revenues, or a movie might not have made any revenue (it got scraped in production). The concern is whether or not to include zero revenue films into consideration, or, if not, which films are removed?

References

https://www.kaggle.com/code/muhammetgamal5/tmdb-5000-movies/notebook