DSC 530

Professor Metzger

Skylar Peven

03-02-2025

**Top 100 Anime EDA**

In the last 10 years anime has exploded in popularity. No longer a niche animation style only enjoyed by Japanese nationals, on obscure websites and adult swim at specific times. Anime is expected to grow at a rate of nearly 10% yearly for the foreseeable future. However, that doesn’t mean every anime made is good. This got me wondering, what variables can lead to an anime being highly rated?

I primarily focused on exploring what variables are correlated with ratings and explored the statistical question of whether a bigger production budget leads to better ratings for the top 100 highest-rated anime. I hypothesized that there is a positive correlation between production budget and anime ratings, suggesting that higher budget anime tend to receive better ratings. However, the analysis indicates a weak correlation between budget and ratings, as well as between release year and budget. My least squares regression model was not statistically significant, and the explanatory variables did not explain much of the variance in the ratings. This suggests that other factors may influence the ratings of the top 100 anime.

One limitation was the small size of the dataset, which comprises only the top 100 highest-rated anime. A larger, more diverse dataset could have allowed for more detailed analysis. I also could have expanded my analysis by including categorical variables in my analyses. Additionally, the analysis did not consider other potential factors that may influence anime ratings, such as genre, popularity, and cultural background. These factors could provide a more comprehensive understanding of the variables that affect anime ratings. I think it would have been interesting to separate the data out by genre and then look for trends.

I completed my analysis as if the variables were normally distributed, but the histograms did not look to be normally distributed, or even bimodal. I wasn’t sure what to do about this, so I left them as is. The analysis also assumed a linear relationship between the variables, but this may not be the case. There may be a non-linear relationship, and I would have needed to use additional modeling techniques and feature engineering to tease them out of the dataset.

One of the main challenges was gathering reliable data on anime production budgets, as this information is not always publicly available. My dataset only contained the top 100 anime, and I have some serious concerns about the accuracy of the data provided. Some of the categorical variables seemed to have incorrect information. I tried to avoid using these in my analysis.

One thing I don’t fully understand is what assumptions I am making with each type of analysis. These things are easy to look up but when the assumptions shouldn’t be made, I’m not quite sure what to do. In some instances, it makes sense to use another type of analysis instead, but I don’t always know what the alternative is. In the anime data most variables did not appear to have a normal distribution which is primarily assumption in many analyses. I was not sure how to handle that better.