What Drives Anime Ratings?

BA222 F1 TEAM 12
Final Project
Professor Munshi
April 23, 2025

Shanaya Mehra	U38530155
Sindhuja Kumar	U93943011
Songying Lyu	U56230446
Gabriella Pranadi	U86976464

Introduction

Why do some anime become instant fan favorites, while others barely make a ripple? With countless titles vying for attention each year, anime studios must ask: what makes a show stand out? Is it the genre, the number of episodes, or the target demographic? This analysis explores how factors like streaming type, genre, theme, studio, source material, demographics, and the number and duration of episodes influence anime ratings using multivariate regression and exploratory data analysis. By identifying key factors that predict high ratings, the goal is to provide production companies and studios with data-driven insights to guide investment and promotion decisions, ultimately understanding how content and popularity drive success.

Data Description

The original dataset contained 15,000 anime entries and 24 variables- 8 numerical (e.g., score, members, favorites, episodes) and 14 categorical (e.g., genre, type, source, studio). After data cleaning, the final dataset included 12,232 entries and 41 variables, comprising 38 dummy variables from one-hot encoding and 3 numerical variables. For the demographics and type columns, one-hot encoding was applied to all unique values due to their limited number of categories. For the genres, themes, source, and studios columns, the top 5 categories were retained, with all other less frequent categories grouped as "Other." This was achieved through the `create_top_n_dummies` function, which uses the `value_counts()` method to determine the most frequent categories and applies a lambda function to categorize values accordingly. This method reduces the dimensionality of the dataset while retaining the most meaningful categorical information. However, this approach can lead to a loss of granularity and potentially obscure meaningful patterns in the data. The dependent variable in this analysis is the weighted_score, which combines the average score (score) and the number of users who rated the anime (scored_by). This technique gives more weight to scores based on a larger number of ratings, using a formula inspired by IMDb's weighted average.

$$WeightedScore(WS) = \frac{(v * R) + (m * C)}{v + m}$$

The above formula incorporates the average score (R), the number of ratings (v), a baseline average score (C), and a threshold for significant votes (m), set at the 75th percentile to reflect a reasonable level of user participation. This ensures the weighted score stays within the original range of 1 to 10 without requiring additional scaling. The independent variables considered include the categorical variables such as genre, themes, demographics, type, studios, and source, along with numerical variables like the number of episodes and duration of each episode.

Missing values were a significant challenge for the demographic, genres, and themes variables. To address this, two key strategies were employed: filling missing demographics based on ratings and imputing missing genres or themes using the synopsis. For missing demographic data, a predefined mapping links anime ratings (e.g., 'PG-13' or 'R') to demographics like 'Shounen' or 'Seinen.' If a demographic is missing but the rating is available, the value is filled based on this mapping, ensuring consistency with the intended audience. For missing genre and theme data, a TF-IDF (Term Frequency-Inverse Document Frequency) approach analyzes the synopsis text to identify keywords most related to the missing values. By emphasizing words that are frequent in a specific anime's synopsis, the method predicts and fills missing genres or themes with contextually relevant terms. This process resulted in imputing over 10,535 missing

values, helping to preserve the dataset's integrity and ensure more accurate analysis and modeling. Columns with high percentages of missing values were dropped as needed.

Following the cleaning process, we began our exploratory data analysis by examining the distribution of the dependent variable. Figure 1.1 presents a histogram of weighted_score, which displays a slightly right-skewed distribution ranging from approximately 5.9 to 9.2. The mean score is 6.91, while the highest frequency of ratings occurs just below this value, indicating a central tendency around 6.8 to 6.9. The distribution has a long tail to the right, suggesting that while most anime are moderately rated, a small number achieve exceptionally high scores.

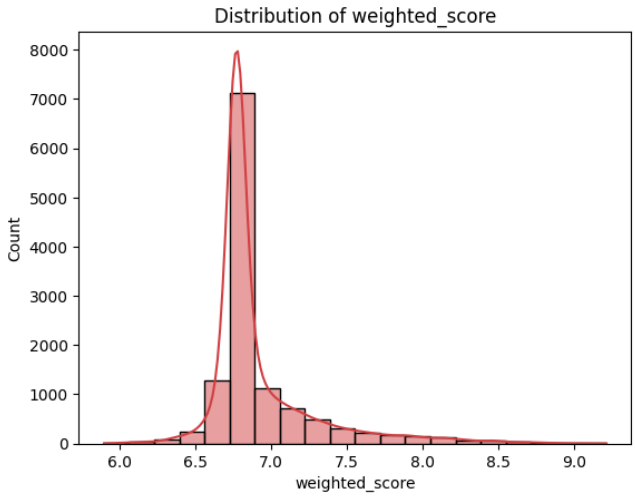

Figure 1.1 Histogram of Weighted Scores

Next, we explored the distribution and relationship of the independent variables with the dependent variable of weighted score. For the numerical variables duration_minutes and episodes, we used scatterplots and correlation matrices to assess potential trends. These graphs revealed little to no linear correlation with the target variable, as shown in Figures 1.2 and 1.3, although a few short-form and long-running anime appeared among the top-rated titles.

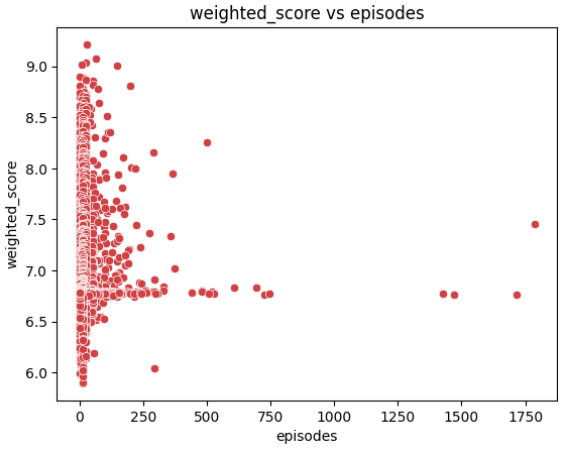

Figure 1.2 Scatterplot of Weighted Scores vs Episodes

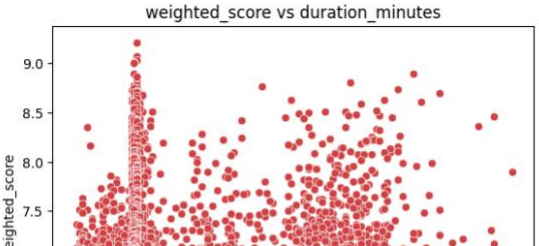

Figure 1.3 Scatterplot of Weighted Scores vs Duration

Next, to quantify the impact of categorical attributes on anime ratings, we calculated the correlation between each independent dummy variable and weighted score. Figure 1.4 displays the top ten dummy variables with the highest positive correlations. Variables like type_TV and source_Manga show the strongest associations with higher scores, followed by source_Light novel, type_Movie, and production studios such as Madhouse. While the correlations are relatively modest in magnitude, they offer useful directional insights into which content types and sources tend to be better received. These findings helped guide our feature selection process during regression modeling.

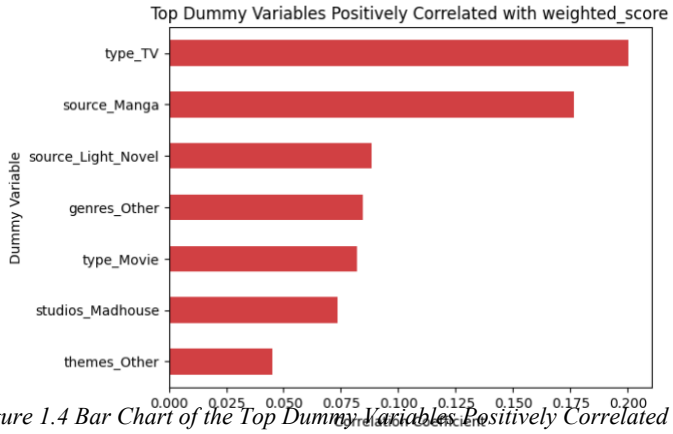

Figure 1.4 Bar Chart of the Top Dummy Variables Positively Correlated with Weighted Score

To complement this, Figure 1.5 below is a heatmap showing the correlation coefficients between various variables in the dataset. The colors represent the strength and direction of the correlations, with red indicating positive correlations and blue indicating negative correlations. Notable strong positive correlations include genres_Hentai with demographics_Josei (0.84) and type_Movie with duration_minutes (0.70). Negative correlations are also present, such as genres_Other with genres_Comedy (-0.53), indicating an inverse relationship. The weighted_score shows only one moderate positive correlation with type_TV (0.20). The matrix provides a comprehensive view of how the different variables are related, offering insights into confounding variables and their potential influence on the weighted score.

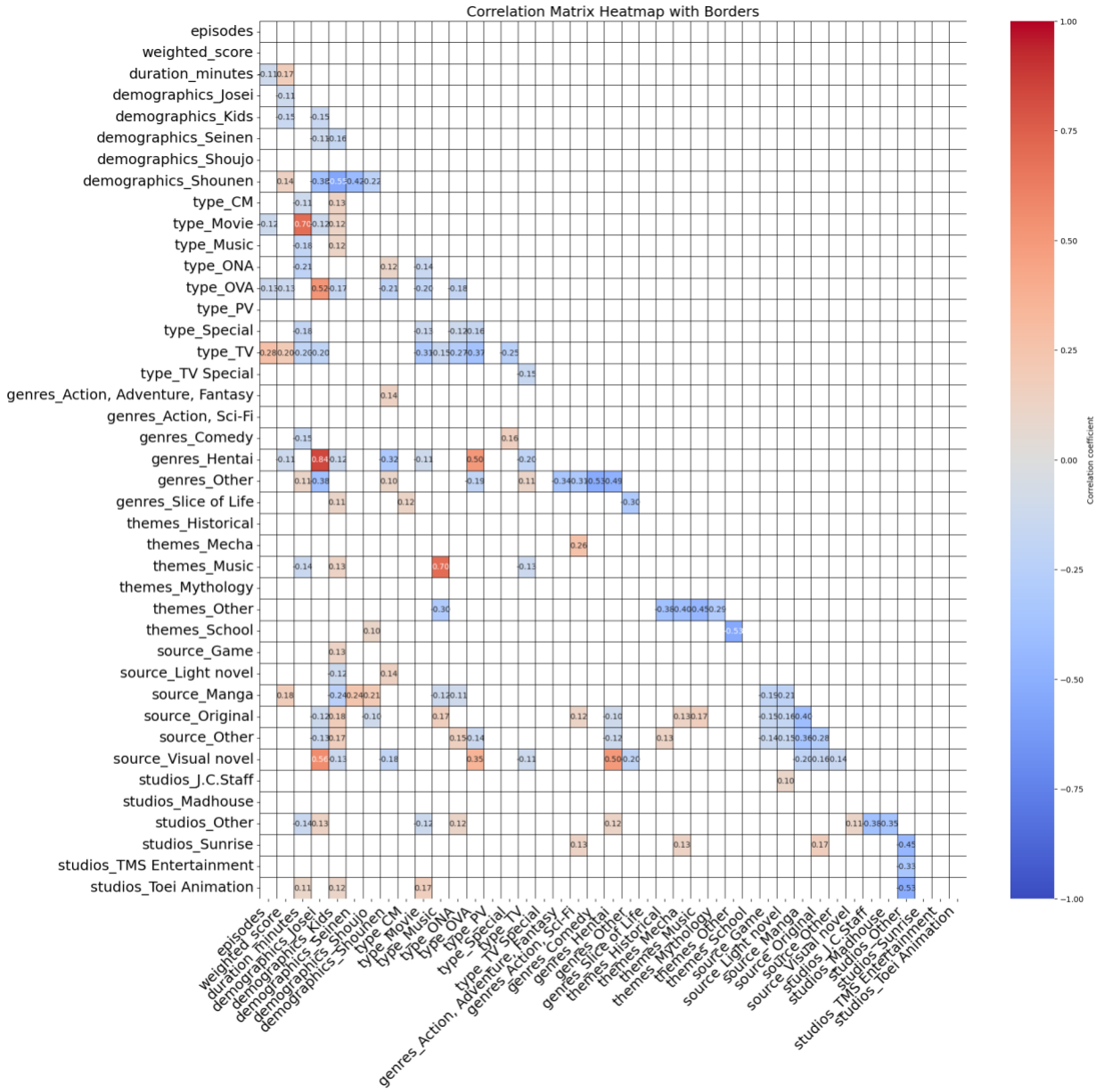

Figure 1.5 Correlation Matrix Heatmap with All Variables

To assess potential confounders in the relationship between anime TV type (type_TV) and weighted score (weighted_score), we compared a baseline regression model (without confounders) to models including various potential confounders, such as the dummy variable across demographics, genres, themes, source types, and studios. A confounder was deemed significant if its inclusion caused a substantial change (greater than 0.001) in the coefficient for type_TV, indicating its influence on the relationship.

Omitted variable bias can occur if relevant confounders are excluded from the model, leading to misleading estimates of the relationship between the variables. By identifying and incorporating these confounders, we reduce the risk of omitted variable bias, ensuring that the model more accurately reflects the true effect of anime TV type on the weighted score.

Through this analysis, we were able to identify 31 confounding variables, including factors such as duration_minutes, demographics_Josei, and demographics_Seinen, that significantly impacted the relationship between anime TV type and weighted score. These confounders will be included in forward selection to develop the final multiple regression model, providing accurate insights into the effect of anime TV type while controlling for other influential factors and minimizing potential bias.

Causal Analysis

In Table 1.1, three regression models were compared to assess the relationship between anime TV type (type_TV) and weighted score (weighted_score), controlling for confounding variables. The univariate regression model showed a positive relationship with a beta coefficient of 0.1473 and a standard error of 0.007, but the adjusted R-squared value of 0.040 indicated a weak explanatory power. The multivariate regression model, which included 31 confounders, resulted in a higher beta coefficient of 0.1901 but also had a larger standard error of 0.016 and an adjusted R-squared of 0.149, indicating better explanatory power at the cost of increased complexity and reduced precision. The multivariate regression model with a few confounding variables, which included 13 confounders, struck a balance between accuracy and simplicity, producing a beta coefficient of 0.1486 with a smaller standard error of 0.008 and an adjusted R-squared of 0.139. This model provided a good balance between capturing the relationship and avoiding overfitting.

	Univariate Regression Model Between Weighted Score and Anime TV Type	Multivariate Regression Model Controlling for All Confounding Variables	Multivariate Regression Model Controlling for A Few Confounding Variables
Beta Coefficient of the Anime TV Type	0.1473	0.1901	**0.1486**
Standard Error	0.007	0.016	**0.008**
Sample Size (Number of Observations)	12232	12232	**12232**
Number of Confounding Variables	0	31	**13**
Adjusted R-Squared	0.040	0.149	**0.139**

Table 1.1 Comparison Among Three Different Regression Models

Our best model, which controlled for 13 confounding variables, had the beta coefficient for type_TV equal to 0.1486. This indicates that, on average, the presence of TV-type anime increases the weighted score by 0.1486 while controlling for other variables. In comparison, the univariate model had a coefficient of 0.1473, showing that not controlling for confounders resulted in a slightly lower estimate. This difference highlights the impact of confounding variables, suggesting that ignoring them led to a more limited understanding of the relationship between TV type anime and weighted score.

Regression Model and Results

Having identified the key variables and potential confounders that may influence anime ratings, we proceeded to quantify their effects using regression modeling. Regression analysis allows us to assess the strength and direction of these relationships while controlling for other factors.

To explore the factors influencing anime ratings, we ran a multivariate regression model using weighted_score as the dependent variable. The model incorporated a range of content-related and demographic factors, as well as other key characteristics, to assess their relationship with anime ratings while controlling for confounding variables. We employed forward selection as our feature selection method due to the high number of one-hot encoded categorical variables in our dataset. Many of these variables are sparse or weakly related to the target, so forward selection allowed us to build a more focused model by starting with no predictors and incrementally adding only those that significantly improved the adjusted R-squared. This approach was especially relevant in our context because it helped reduce overfitting, improved interpretability, and ensured that only the most impactful features were included in the final regression model.

The resulting model yielded an adjusted R-squared value of 0.139, which indicates that our model explains approximately 14% of the variance in the weighted score. While this suggests a moderate level of explanatory power, there are likely other factors, such as production values or international appeal, that contribute to anime ratings but are not captured by this model or the dataset. Although this is a moderate level of explanatory power, it suggests that the factors we included have a significant, though not overwhelming, influence on the ratings.

The results in Table 1.2 highlight several factors that significantly impact anime ratings. TV-type anime, with a coefficient of 0.1486, tends to receive higher weighted scores compared to non-TV types, and this relationship is highly significant with a p-value of 0.000. This suggests that TV-type anime consistently perform better in terms of ratings than other types of anime when all other factors are held constant. The source of anime also plays a crucial role in determining ratings. Anime based on Manga (coefficient = 1.0538) and Light Novels (coefficient = 1.0439) shows particularly high weighted scores, both with very low p-values (0.000), indicating strong positive relationships with ratings. These findings suggest that adaptations from Manga and Light Novels are strongly associated with higher ratings, possibly due to a built-in fanbase or higher production value.

In addition to the source, demographics and studios also significantly influence ratings. Anime targeted at the Shounen demographic (coefficient = 0.0761) tends to receive higher ratings, as does anime produced by Madhouse (coefficient = 0.1192), with both relationships being highly significant (p-values of 0.000). These results indicate that both the target audience and the studio behind the anime play important roles in its reception. The duration of episodes also shows a small but statistically significant positive effect on ratings (coefficient = 0.0027), suggesting that longer episodes are slightly associated with higher ratings. While the effect is small, it is still notable in the context of the model.

	Coefficient	P-value
Intercept	5.7561	0.000

type_TV[T.True]	0.1486	0.000
source_Manga[T.True]	1.0538	0.000
demographics_Shounen[T.True]	0.0761	0.000
source_Light_Novel[T.True]	1.0439	0.000
studios_Madhouse[T.True]	0.1192	0.000
source_Game[T.True]	0.8651	0.000
type_OVA[T.True]	-0.0388	0.000
source_Other[T.True]	0.9128	0.000
type_ONA[T.True]	0.0197	**0.072**
source_Original[T.True]	0.9269	0.000
source_Visual_Novel[T.True]	0.9535	0.000
genres_Other[T.True]	0.0135	**0.066**
duration_minutes	0.0027	0.000

Table 1.2 Final Multivariate Regression Model Summary

On the other hand, the variable genres_Other has a minimal positive effect on ratings (coefficient = 0.0135), with a marginal p-value of 0.066, indicating that its effect is less significant compared to other variables. Similarly, anime type ONA has a positive coefficient of 0.0197, but its p-value of 0.072 indicates a weaker relationship with ratings, suggesting that the effect of being an ONA-type anime is not as strong or reliable. Lastly, anime type OVA has a negative coefficient of -0.0388, suggesting a slight decrease in ratings for OVA-type anime. However, this effect is statistically significant with a p-value of 0.000, indicating that, while the effect is small, it is robust across the dataset.

Overall, the model as a whole is statistically significant, with an F-statistic of 165.4 and a p-value of 0.00, indicating that the independent variables collectively explain a meaningful portion of the variance in the weighted score. This highlights the importance of content-related and demographic factors in influencing anime ratings.

Conclusion

This analysis provides valuable insights into the factors that influence anime ratings, focusing on variables such as streaming type, genre, theme, studio, source material, demographics, and episode details. The results reveal that TV-type anime, Manga, and Light Novel adaptations, along with the Shounen demographic, have the strongest positive impact on ratings. TV-type anime, in particular, shows the most significant relationship with higher ratings, while adaptations from Manga and Light Novels tend to yield particularly high weighted scores. These findings offer critical guidance to production companies and studios, helping them identify which anime characteristics to prioritize for investment and promotion.

Understanding these factors enables companies to make more informed decisions that can increase the chances of success in a highly competitive market.

The methodology employed in this analysis- using multivariate regression and exploratory data analysis- enabled a comprehensive examination of the relationship between various factors while accounting for confounders. A key component of the analysis was the use of the weighted score, a metric that combines both the average score and the number of ratings. This allowed us to account for the influence of highly-rated, well-reviewed shows with larger fan bases, providing a more accurate representation of a show's success than using raw average scores alone. Additionally, we employed TF-IDF to impute missing genres and themes, leveraging the text-based nature of anime synopsis. By focusing on recurring terms most strongly associated with specific genres or themes, this method provided an effective way to address missing data. However, alternative methods, such as KNN imputation, could be explored in future analyses to further enhance accuracy and fill in gaps in the dataset.

Despite the robustness of the methodology, this analysis is limited by the reliance on observational data, which can reveal correlations but cannot establish causality. The findings, therefore, should be interpreted with caution- while we can confidently identify relationships between variables, we cannot definitively say that one factor causes another. Additionally, the model's explanatory power, as indicated by an adjusted R-squared value of 0.139, suggests that there are other influential factors not captured by the model. While the model explains some variance in anime ratings, a large portion remains unexplained, highlighting the complexity of the factors that drive ratings.

To address these limitations and refine the analysis, future work could incorporate more granular data. For example, including episode-level details and specific viewer demographics could provide deeper insights into how these factors contribute to the overall ratings. Furthermore, employing more advanced modeling techniques, such as machine learning algorithms like decision trees, random forests, or neural networks, could help capture non-linear relationships between variables and improve predictive power. These methods could also uncover interactions between variables that linear regression may not fully capture. Additionally, sentiment analysis on anime reviews, considering both textual sentiment and viewer engagement, could provide a richer understanding of viewer preferences and emotional reactions, which are likely significant contributors to ratings.

In conclusion, this analysis has provided key insights into the factors that drive anime ratings, offering production studios actionable data to optimize their strategies for better viewer engagement and success. By identifying the characteristics that predict high ratings, studios can make more informed decisions about which types of anime to produce, promote, and invest in. Future research should build upon this foundation by incorporating more detailed data, applying advanced predictive models, and exploring new avenues for understanding viewer sentiment and engagement.