**Into the World of Anime: An Exploration and Analysis**

**Crystal Kao, Pranideep Meka, Swathi Murali Srinivasan**

**Introduction**

This report delves into a comprehensive analysis of anime popularity in 2022, utilizing data extracted from Anime Planet's 'Anime DataSet 2022.' With over 18,000 anime entries encompassing 16 variables each, our exploration aims to unravel the nuanced factors shaping the landscape of anime viewership.

As we scrutinize the dataset, our primary objectives are to predict user ratings based on critical attributes such as studio, release season, and tags. Additionally, we aim to identify the key determinants influencing an anime's rating and construct a predictive model for overall popularity.

Our findings provide actionable insights for industry stakeholders, including studio executives and content creators, offering a strategic lens to optimize content creation and anticipate trends within the dynamic realm of anime. This executive summary serves as a prelude to a detailed report that navigates the intricacies of our analysis and presents a roadmap for leveraging data-driven perspectives in the evolving world of anime entertainment.

**Dataset**

The dataset this analysis uses is sourced from Kaggle, a platform that often hosts data science competitions and contains thousands of datasets. The specific dataset used is called Anime DataSet 2022 and contains web scraped data from Anime Planet (Mane, 2022). Anime Planet is a place where you can “discover anime and manga, track your progress, watch anime, read manga” (Anime Planet, n.d.). There are over 18,000 anime entries and 17 factors available that will be explored in the following analysis.

**SMART (Specific, Measurable, Attainable, Relevant, Timely) Questions**

The goal of our analysis is to answer the following SMART questions:

1. How does the type of anime (e.g., TV series, movie, OVA) relate to the average user ratings on Anime Planet? Are there significant differences in ratings based on the type of anime?
2. Can we predict the user ratings of anime based on the available information such as the studio, release season, and tags? What factors have the most significant impact on an anime's rating, and can we build a predictive model for anime popularity?

To do this we begin with exploring and cleaning the data.

**Methodology**

A screenshot of a computer

Description automatically generatedAt a glance, our data contains 17 columns with 18,495 entries available.

*Fig. 1: Summary of column names and number of entries*

A screen shot of a computer

Description automatically generatedHowever, if we dive deeper, we can see that many of the columns contain large amounts of missing values.

*Fig. 2: Number of missing values in each column*

The easiest way to clean the data is to drop all missing values. However, this method is not a good one due to the fact that there are too many missing values and blindly deleting all of them results in having a dataset with only 40 entries.

A screenshot of a computer

Description automatically generated

*Fig 3. Summary of dataset after removal of all missing values.*

Since that was not a viable option, we returned back to the original dataset and cleaned it with other methods. First new columns were created to have values based on integers instead of strings. ‘’Voice\_actors’, ‘staff’, ‘Related\_anime’, ‘Related\_Mange’, and ‘Tags’ were all converted to columns by count for the number of people, anime, manga, and tags. ‘Related\_anime’ and ‘Related\_Mange’ counts were also added together to create a ‘rel\_media\_count’ column. Unnecessary spaces were removed from the ‘Type’ column. ‘Japanese\_name’, ‘End\_year’, ‘Content\_Warning’ and ‘Related\_Mange’ were then all removed from our dataset due to more than 50% of the columns having missing values and are not relevant to our current analysis. With these changes, the number of missing values went down drastically compared to what we started with.

A screenshot of a computer

Description automatically generated

*Fig. 4: Number of missing values in each column after cleaning.*

With this, the dataset was deemed clean enough to do some exploratory data analysis before doing more relevant cleaning and feature engineering to answer our SMART questions.

**Exploratory Data Analysis**

A graph of a number of points

Description automatically generated with medium confidenceOne of the largest components of using Anime Planet and why collecting data on this platform is useful is the rating system. To get a good picture of how anime are doing in terms of rating, the following graph was created.

*Fig. 5: Histogram of count of anime by ratings.*

The majority of anime tend to fall within the rating range of 3 to 4, indicating that only a small number of anime received exceptionally high ratings. Interestingly, the data follows a normal distribution which is relatively rare for real world data.

Next, to get a sense of the range of data, a graph of the year range of anime was created.

A graph of a number of years

Description automatically generated with medium confidence

*Fig. 6: Histogram of count of anime by release year*

The data contains data on anime spanning over 100 years, basically back to the beginning of when the genre was first created. Most anime were released within the last 10 to 15 years, which aligns with the significant growth observed in the manga and anime industry during the late 2010s (Anime Industry Data, n.d.).

A graph of a bar chart

Description automatically generated with medium confidence

*Fig. 7: Bar chart of anime count by media type*

Since one of our SMART questions focuses on media type, we took a preliminary look at the amount of anime released based on media type. TV shows are the most common form anime is released as, followed by movies.

Another interesting field available in the data is anime studios. As seen below, out of the top 15 anime studios, Toei Animation produces the most anime of any studio and the gap between first and second place is over 200 anime.

A graph of a number of movies

Description automatically generated with medium confidence

*Fig. 8: Bar chart of top 15 anime studios by count of anime*

Interestingly there's a Chinese animation studio in the top 15 animation studio list as well called Shanghai Animation Film Studio. This is because Chinese anime, also known as donghua, has also become popular in recent years and is often tracked on anime platforms due to the genre’s similarities. Many donghua are also outsourced to Japanese studios.

Another field of interest is the number of episodes. Most anime movies are counted as 1 episode which makes sense in the bar graph as to why there are so many at the 1 episode mark. The next most common episode count is around 12, which is the average length for TV shows or Original Video Animations (OVA).

A graph of a number of episodes

Description automatically generated

*Fig. 9: Histogram of number of episodes by anime count*

To slightly increase the complexity of the exploratory analyses, we next looked at anime ratings by media type.

A chart of a variety of levels

Description automatically generated with medium confidence

*Fig. 10: Box plots of anime ratings by media type*

TV shows have the highest average rating compared to all other media types, with TV Sp (TV Specials) at a close second place.

A chart with different colored dots

Description automatically generatedAnother way to look at the media type is compared to studio data.

*Fig. 11: Scatterplot of anime media types released by the top 15 studios.*

Most studios mostly create TV shows (brown dots) and movies (orange dots).

Another factor to explore is the release season. Anime releases are often categorized by season such as summer, winter, fall, and spring and are often talked about as easy groupings in the anime-watching community by discussing if the coming fall season anime will be good or not.

A chart with different colored squares

Description automatically generated

*Fig. 12: Box plots of anime ratings by release season.*

When looking at ratings by season, there interestingly does not seem to be much of a difference between the seasons as all of them have a mean rating of ~3.5.

A screenshot of a computer

Description automatically generatedWe also took a look at a correlation matrix for the fields we are interested in.

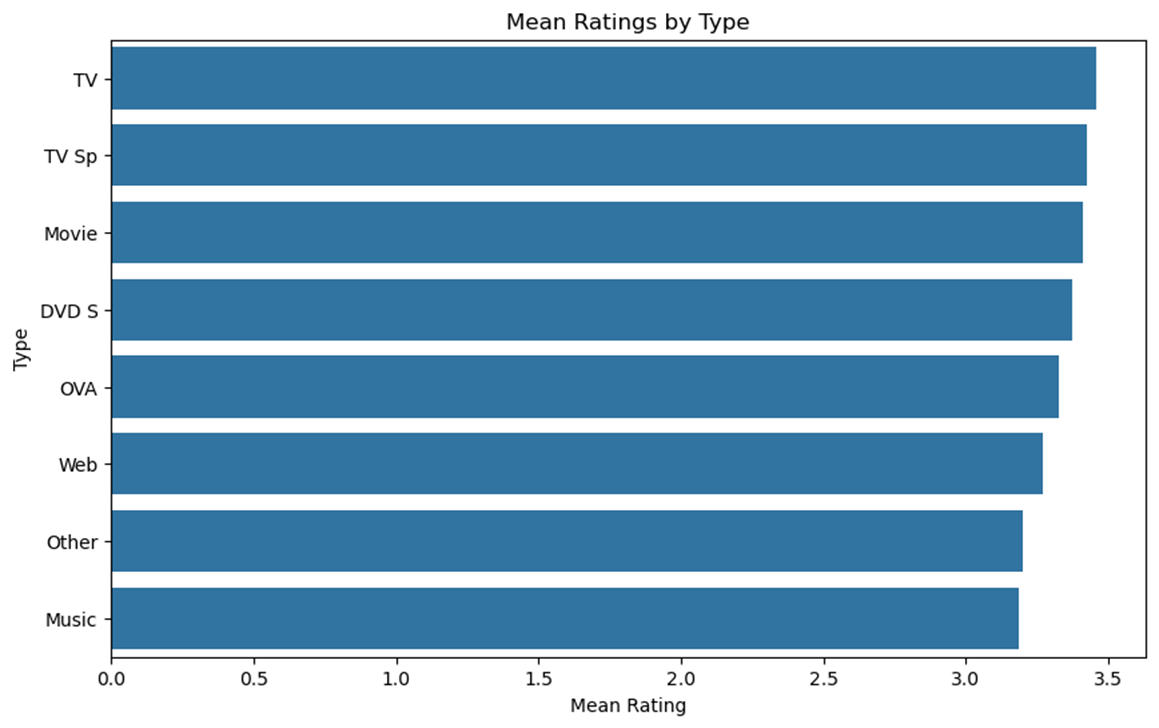
*Fig. 13: Correlation matrix of the dataset*

Before interpreting the correlations, there are some caveats to the data. Rankings are created based on ratings so ignore that correlation (black -0.95). The correlation between related media and related anime (beige 0.94) and manga (orange 0.77) can also be ignored because those fields make up related media.

From the matrix, there is an interestingly strong negative correlation between the number of voice actors and rank. The count of staff and voice actors have a relatively high correlation with each other. This makes sense since the more voice actors you have, the more staff you need to support them.

**Results**

**SMART Question 1: How does the type of anime (e.g., TV series, movie, OVA) relate to the average user ratings on Anime Planet? Are there significant differences in ratings based on the type of anime?**

****

*Fig. 14: Bar chart of mean anime ratings by media type*

**A screenshot of a computer

Description automatically generated**

*Fig. 15: Table of mean anime ratings by media type*

Based on figures 14 and 15, there is not much of a difference between anime ratings depending on their media type. The averages were all within 3.1-3.5 with TV in the lead but there are no significant differences. This can also be seen in the box plots in figure 10.

**SMART Question 2: Can we predict the user ratings of anime based on the available information such as the studio, release season, and tags? What factors have the most significant impact on an anime's rating, and can we build a predictive model for anime popularity?**

To answer SMART question 2, some further cleaning of the data was required before building a regression model.

Missing episode values were filled using the median, which was 12 episodes. This makes most sense to us as previously discussed, as anime seasons typically have 12 episodes per season. Missing release years were also filled using the median. Missing studio values were filled by randomly selecting a studio. The ‘Tags’, ‘Studio’, ‘Type’, and ‘Release\_season’ were transformed from a column to a list to facilitate binary classification through one-hot encoding.

A screen shot of a computer

Description automatically generated

*Fig. 16: Example of tags available in the dataset*

**Data Preparation**

We begin by creating a dedicated DataFrame, predicting\_df\_clean, which comprises rows with unavailable ratings. This dataset serves as the basis for predicting ratings in the final stage. To ensure meaningful analysis, irrelevant columns such as "Rating," "Name," "Rank," "Tags," and "index" are dropped.

**Train-Test Split**

To evaluate the performance of our predictive model, we divided the available data into two sets: a training set (80% of the data) and a testing set (20% of the data). This division allows us to train the model on one subset and assess its predictive accuracy on another.

**Model Building and Evaluation**

For the predictive model, we opted for an XGBoost Regressor, a powerful algorithm known for its ability to handle intricate relationships within the data. After training the model on the training set, we evaluated its performance using key metrics:

A screenshot of a computer program

Description automatically generated

*Fig. 17: Model evaluation metrics*

* The MAE was calculated to be 0.2231, representing the average absolute difference between the predicted and actual ratings. This metric provides a clear indication of the model's accuracy.
* The MSE was found to be 0.0871, quantifying the average squared difference between the predicted and actual ratings. It offers insights into the spread of errors.
* The R² value was determined to be 0.1686, indicating the proportion of variance in the dependent variable (ratings) that can be explained by the independent variables (features). A higher R² suggests a better fit of the model to the data.

**Result Comparison**

To further analyze the model's effectiveness, we compared specific examples of actual and predicted ratings. The following represents a subset of the comparison:

**Example 1:**

* Predicted Rating: 3.04
* Actual Rating: 2.59

**Example 2:**

* Predicted Rating: 3.57
* Actual Rating: 3.82

**Example 3:**

* Predicted Rating: 2.85
* Actual Rating: 2.86

**Example 4:**

* Predicted Rating: 3.04
* Actual Rating: 3.04

**Example 5:**

* Predicted Rating: 3.42
* Actual Rating: 3.44

These examples provide a snapshot of the model's predictive performance. Comparing predicted and actual ratings for individual cases allows for a more granular understanding of where the model excels and where it may have limitations. In these instances, we observe varying degrees of closeness between predicted and actual ratings, indicating the model's ability to capture the nuances of anime popularity. Further exploration of a larger set of examples can provide more comprehensive insights into the model's overall performance.

**Predictions for Unavailable Ratings**

After successfully training and evaluating the model, we applied it to predict ratings for anime entries that initially had unavailable ratings. This step extends the model's utility beyond the training and testing sets, allowing us to make predictions for real-world scenarios.

A screenshot of a black screen

Description automatically generatedA screenshot of a cell phone

Description automatically generated

*Fig. 18: Predictions Fig. 19: Comparison of actual and predicted values.*

**Visualizing Results**

To aid in the interpretation of the model's performance, a histogram was plotted. This visualization provided an overview of the distribution of differences between actual and predicted ratings. It's a helpful tool for understanding the overall accuracy and any potential biases in the model.

A graph of a normal distribution with Ryugyong Hotel in the background

Description automatically generated

*Fig. 20: Comparison of actual and predicted values.*

**Conclusion**

In conclusion, our comprehensive analysis and the developed predictive model have shed light on the questions central to our investigation:

**Can we predict the user ratings of anime based on available information such as the studio, release season, and tags?**

* The results of our modeling efforts suggest a positive answer to this question. The XGBoost Regressor, trained on a dataset comprising key features such as studio, release season, and tags, demonstrated an ability to make reasonably accurate predictions of user ratings. The model's predictive performance, as indicated by metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²), provides confidence in its capability to generalize from the training set to predict ratings for anime entries in the testing set.

**What factors have the most significant impact on an anime's rating, and can we build a predictive model for anime popularity?**

* Through our analysis, we identified the most influential factors affecting an anime's rating. The model, leveraging features such as studio, release season, and tags, highlighted their significance in determining anime popularity. The Result Comparison section further illustrated the model's ability to capture nuances, as evidenced by the closeness of predicted and actual ratings in specific instances.

**The Impact of Our Analysis and Model:**

Our analysis and the resulting model provide actionable insights for various stakeholders in the anime industry. Studio executives, content creators, and distributors can leverage this predictive model to anticipate the potential popularity of upcoming anime releases. Understanding the factors that significantly contribute to an anime's rating allows for strategic decision-making, such as targeted marketing efforts or content adjustments to align with audience preferences.

In essence, our analysis not only answers the posed questions but also equips industry professionals with a valuable tool for informed decision-making, contributing to the optimization of anime content creation and distribution. As the industry evolves, our model provides a forward-looking perspective on anime popularity, aligning with the dynamic nature of audience preferences and contributing to the growth and success of the anime landscape.

**References**

*Anime Industry Data: 日本動画協会*. The Association of Japanese Animations. (n.d.). https://aja.gr.jp/english/japan-anime-data

Mane, V. (2022, January 16). *Anime dataset 2022*. Kaggle. https://www.kaggle.com/datasets/vishalmane10/anime-dataset-2022/data

*Welcome to anime-planet*. Anime. (n.d.). https://www.anime-planet.com/