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

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

**Introduction**

**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**

At a glance, our data contains 17 columns with 18,495 entries available.

A screenshot of a computer

Description automatically generated

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

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

A screen shot of a computer

Description automatically generated

*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**

One 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.

A graph of a number of points

Description automatically generated with medium confidence

*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.

Another way to look at the media type is compared to studio data.

A chart with different colored dots

Description automatically generated

*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.

We also took a look at a correlation matrix for the fields we are interested in.

A screenshot of a computer

Description automatically generated

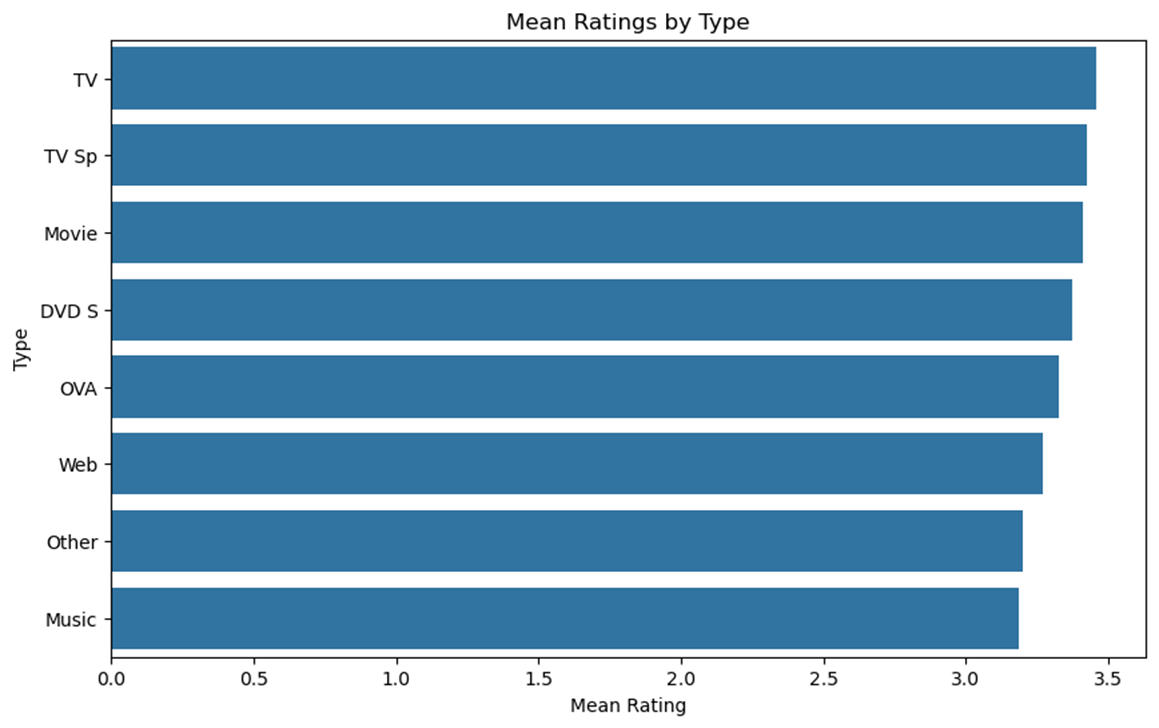
*Fig. 13: Correlation matrix of the dataset*

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*

**Conclusion**

**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/