The nature that determine an anime show's popularity involves the source material of the adaptation and the years it was released.*

Justin Abando

21 April 2023

There are online lists and forums where users can vote on what they believe to be the most popular anime such as MyAnimeList (MAL). However, there are no metrics that can determine why specific anime stand out, and so data was retrieved to determine what factors could be responsible for a show's popularity. It was found that the source material and year of the anime's release were significant in determining how many users of MAL watched the anime. The findings of this paper suggests reasons why a particular anime becomes popular, which can be relevant to those that are new to watching anime or those that want something new to watch as they would gravitate to shows that receive traction by a broader demographic.

1 Introduction

Japanese animation (referred hereafter as anime) is a medium that has been exploding in popularity in recent decades. With an already large catalogue of shows to choose from, many newcomers to the medium seek guidance in what shows they should start watching as their entry to anime. One such website that helps with this is MyAnimeList (MAL), which serves as both a social networking platform for anime fans, but also as a social cataloguing website to keep track, rank, and review the shows that you can watch. MAL is one of the most frequently used anime catalogues online, with millions of active users and website visits every month.

With a large user base, MAL members can personally rate and review the shows that they watch and catalogue them in their anime lists. While this is useful at the individual level to keep track of their own journey and see the depth and breadth of the shows they watch, it is

*Code and data are available at: https://github.com/justabando/malanalysis.

also useful at the community level. The rating and reviews that are given from individuals are aggregated and averaged using all of the data from other MAL users to give an overall subjective ranking of each show in terms of popularity and rating. It should be noted that a show's ranking is not defined by objective means, but is the majority voice of all the subjective reviews left by the users of the website. With MAL, popularity and rating are both metrics used to sort and determine the top anime based on the reviews and ratings that were given to them by users. While the rating of the show matters for popularity to an extent, a large part of it's score is dependent on the content of the show itself. The rating is dependent on a scale of 1-10, and takes the average score of all the users that have rated the show. Meanwhile, popularity as a ranking metric demonstrates how much exposure the show has gotten and its outreach across as many demographics as possible, and is a ranking that is determined by the amount of users that rated the show regardless if it was positive or not.

This presents two ways for newcomers to judge a show before watching. They can be inclined to start a show that is subjectively rated as one of the best as ranked by the MAL community, or choose one that has gained the most traction despite any apparent flaws or drawbacks that may deter a positive watching experience. As such, it should be noted that popularity is not a metric that is inherently positive, but can also imply the notorious status of a show, especially if it is ranked low in terms of rating but manages to be much more popular than its rating score suggests. This creates an interesting method to identify shows that have managed to be popular despite its poor ranking performance. It could be that the show has notoriously bad traits that often gets shared to others to ridicule or prank unsuspecting friends to watch and experience a show that can be "so bad that its good". The anime may have been able to maintain a good quality throughout its run time, but eventually hit a point in the plot that polarizes the show's watchers. Regardless, there are many ways to interpret the collective opinion of a show on MAL with the presence of ranking both its ratings and popularity. While the rating of a show is indicative of how well-received the show is strictly by the subject and content of the show itself, the popularity of a show can be shrouded in mystery and depend on many logistical factors the put an anime in the limelight.

With all of this in mind, the objective of this analysis is to determine the factors that influence an anime's popularity. How an anime becomes well-known and circulated through the community and even seeps through the cracks and becomes mainstream is an interesting topic to discuss due to the many facets of the show in terms of its production. To determine these exact causes, a statistical analysis is required to contextualize the exact factors that are relevant in determining an anime's popularity. These factors will be expanded upon in Section 2.

It was found that the source material of the anime that it was adapted into had a significant impact on MAL viewership of anime, as shown with a linear regression model found in Section 3. Additionally, _____.

The findings of this paper suggest that the popularity of a show is influenced by a variety of factors that may not be obvious by just watching the show itself. These findings can help someone who is new to watching anime, or those that want something new to watch to find something that gained traction within the community. Popularity can be a factor that is more

important than the ranking of the show's quality in many cases, and this is due to the context that the show presents. An anime may be very well-received, but it may be less known within the broader community. With less overall viewership of an anime, the show may still be ranked relatively high in terms of rating due to its niche community that watches the show, which can inflate its overall ranking in its ratings. This is why ratings are not necessarily the best metric to depend on for newcomers to find a show to watch, as a niche show may not be widely accessible to a broad audience. However, popularity ranking can be much more indicative of a potential show someone can watch due to its exposure that is not dependent on the taste or bias of the viewer.

This statistical analysis is divided into several sections, with Section 2 giving context and describing the variables that will be used to determine the underlying reason behind an anime's popularity. Section 3 describes the model used to create the correlation between a show's popularity and the factors that possibly influence it. Section 4 demonstrate the overall findings of the statistical analysis and puts into perspective how these factors affect a show's popularity, and Section 5 covers a discussion of the observations that were made and inferring the causes based on the results of the data.

2 Data

2.1 Data Collection

This report focuses on the data available from MyAnimeList (MAL) since July 2022. MAL is a social networking platform for anime and manga that also doubles as a database for said mediums. The data itself was obtained from Kaggle from Andreu Vall Hernandez, which gathered data from the official MAL API, as well as a third-party Jikan API. A limitation to the way the data was gathered was that it was aggregated from an outside source, however there is currently no way to publicly access the official MAL API without requesting it through an application process as it is currently in active development.

Despite the data from MAL being used in this analysis, there are other anime databases that could have been used, such as Anilist and AniDB, MAL has its prominence within the anime community as exemplified by its high user and visitor counts, making it a more viable option to compare the primary measurement of interest: popularity.

Regardless, data from MAL was imported and analyzed in R (R Core Team 2023). To aid in this analysis, the `dplyr` (Wickham et al. 2023) and `ggplot2` (Wickham 2016) packages from the `tidyverse` package (Wickham et al. 2019) were used to clean the data set and generate graphs to visualize it respectively. The `knitr` (Xie 2021) and `kableExtra` (Zhu 2021) packages were used to create the tables used within this analysis. The table used to generate the summary of the regression analysis was tidied using the `broom` package (Robinson, Hayes, and Couch 2023).

2.2 Data Cleaning and Description

Cleaning involved filtering specific variables within the data set to include that were essential to the analysis. The final data set contains the following variables:

- `title`: the title of the anime.
- `type`: the medium of how the anime is delivered (e.g. tv show, movie, etc.).
- `score`: aggregated score of the anime out of 10, contributed by users of MAL that both have the anime in their list as well as rated it numerically out of 10.
- `scored_by`: the number of MAL users that have the anime in their list and rated the anime out of 10.
- `members`: the number of MAL users that have the anime in their list.
- `episodes`: the number of episodes that a given anime has.
- `source`: the type of source material that the anime was adapted from (e.g. manga, light novel, anime original, etc.).
- `start_year`: the year in which the anime first started to broadcast.

The important variable of note to consider from this list is `members`. This is a quantitative variable that popularity ranking on MAL is completely dependent on. The popularity ranking of an anime is directly affected by how many members an anime has, and so an anime that has the most members would have the #1 popularity ranking. Other than `title`, the rest of the variables will be tested to see if they have any kind of influence on the amount of users that have specific shows on their lists.

2.3 Preliminary Data Analysis

As seen in Table 1, users influence what shows are considered to be the top rated among the MAL community. However, an interesting observation to note is that although the common denominator with the entries of the top 10 shows is the high user ratings, there seems to be a disparity with the amount of users that have simply watched the show compared to the amount of users that have actually rated the show. The top rated show, Fullmetal Alchemist: Brotherhood, has approximately 3,000,000 members, whereas the seventh rated show, Hangyaku no Lelouch R2, has approximately 1,600,000 members, which is almost half of the former. As a result, the popularity of a given anime cannot be determined simply by the aggregated user score, but must take into account how many people actually have the show in their list, regardless if they rated it or not.

Table 1: Sample from the top rated anime of MAL, which includes the variables that MAL users influence.

Rank	Title	User Rating	# of Ratings	# of Members
1	Fullmetal Alchemist: Brotherhood	9.13	1871705	2932347
2	Hunter x Hunter (2011)	9.04	1509622	2418883
3	Shingeki no Kyojin Season 3 Part 2	9.07	1329500	1881734
4	Steins;Gate	9.08	1252286	2269121
5	Koe no Katachi	8.95	1398608	2001335
6	Kimi no Na wa.	8.86	1675677	2392235
7	Code Geass: Hangyaku no Lelouch R2	8.91	1079799	1587851
8	Shingeki no Kyojin: The Final Season	8.83	1080165	1629273
9	Sen to Chihiro no Kamikakushi	8.78	1149660	1633281
10	Code Geass: Hangyaku no Lelouch	8.70	1266726	2013999

Table 2 describes the same top ten shows on MyAnimeList, but from the perspective of the information that describes the anime itself. There are many facets that can differentiate anime, such as the source material of the show or the medium in which it is presented. Source material such as manga and visual novels can provide much more context within a story and be rich with information and world building, which can contribute to an overall more appealing story to users. The medium of which the anime is presented, such as if the anime was a feature film, may be more accessible to a general audience due to a smaller time commitment to watch the show compared to episodic series that require more time to watch and process.

Table 2: Sample from the top rated anime of MAL, which includes the information of each anime itself.

Rank	Title	Show Type	Episode Count	Source Material	Start Year
1	Fullmetal Alchemist: Brotherhood	tv	64	manga	2009
2	Hunter x Hunter (2011)	tv	148	manga	2011
3	Shingeki no Kyojin Season 3 Part 2	tv	10	manga	2019
4	Steins;Gate	tv	24	visual novel	2011
5	Koe no Katachi	movie	1	manga	2016
6	Kimi no Na wa.	movie	1	original	2016
7	Code Geass: Hangyaku no Lelouch R2	tv	25	original	2008
8	Shingeki no Kyojin: The Final Season	tv	16	manga	2021
9	Sen to Chihiro no Kamikakushi	movie	1	original	2001
10	Code Geass: Hangyaku no Lelouch	tv	25	original	2006

The main variables of interest as mentioned previously are medium of the show, episode count, source material, initial release year, and studio. These are all important and can possibly influence the popularity of an anime without giving any actual information of what the anime is about.

3 Model

A linear regression model can be used to predict how each of the qualitative independent factors can affect how many users of MAL watch a given anime. One particular variable of note to be examined is the source material of the show, as there is the consideration of fans of the source work of the anime adaptation. Those that follow the source material would naturally have some level of interest in an anime adaptation, and so this model accounts for how significant the source material is in influencing MAL viewership of a show.

The following linear regression model was created below to illustrate this:

$$\text{MAL Viewership} = \beta_0 + \beta_1 \cdot \text{Source Material} + \epsilon$$

With this model uses the following components:

- β_0: The intercept of the regression line.

- β_1: The coefficient representing the relationship between source material and members of an anime.
- ϵ: The residual or error term.

Table 3 illustrates the results of the linear regression modelling, where it is seen that several popular source material formats had a significantly low p-value, suggesting there is a correlation between the type of source material the anime is adapted from and MAL viewership. Prominent source material formats such as manga, light novels, and original anime all have p-values less than 0.05, suggesting there is a correlation between the two variables.

Table 3: Linear regression model summary statistics of the impact of source material on MAL user viewership

Source Type	Estimate	Std. Error	Test Statistic	P-value
(Intercept)	70233.34	12046.10	5.8304	0.0000
sourcebook	-59092.15	25706.65	-2.2987	0.0215
sourcecard game	-44227.64	28141.57	-1.5716	0.1161
sourcegame	-44085.70	13862.39	-3.1802	0.0015
sourcelight novel	133378.26	13872.91	9.6143	0.0000
sourcemanga	29550.36	12455.45	2.3725	0.0177
sourcemixed media	-32489.54	29186.30	-1.1132	0.2657
sourcemusic	-62542.76	17364.22	-3.6018	0.0003
sourcenovel	-27151.27	14965.25	-1.8143	0.0697
sourceoriginal	-35062.29	12488.01	-2.8077	0.0050
sourceother	-50574.04	15516.33	-3.2594	0.0011
sourcepicture book	-66110.43	32134.59	-2.0573	0.0397
sourceradio	-66546.14	88191.91	-0.7546	0.4505
sourcevisual novel	-34650.81	13454.00	-2.5755	0.0100
sourceweb manga	40200.63	17309.44	2.3225	0.0202
sourceweb novel	-32317.65	55504.63	-0.5823	0.5604

4 Results

4.1 Show Type

To visualize how the medium of the anime affects MAL viewership, Figure 1 demonstrates the influence of format. The most common format is TV anime, followed closely by movie anime. When looking at the MAL members that have watched it, it can be observed that the majority of users watch shows within those formats.

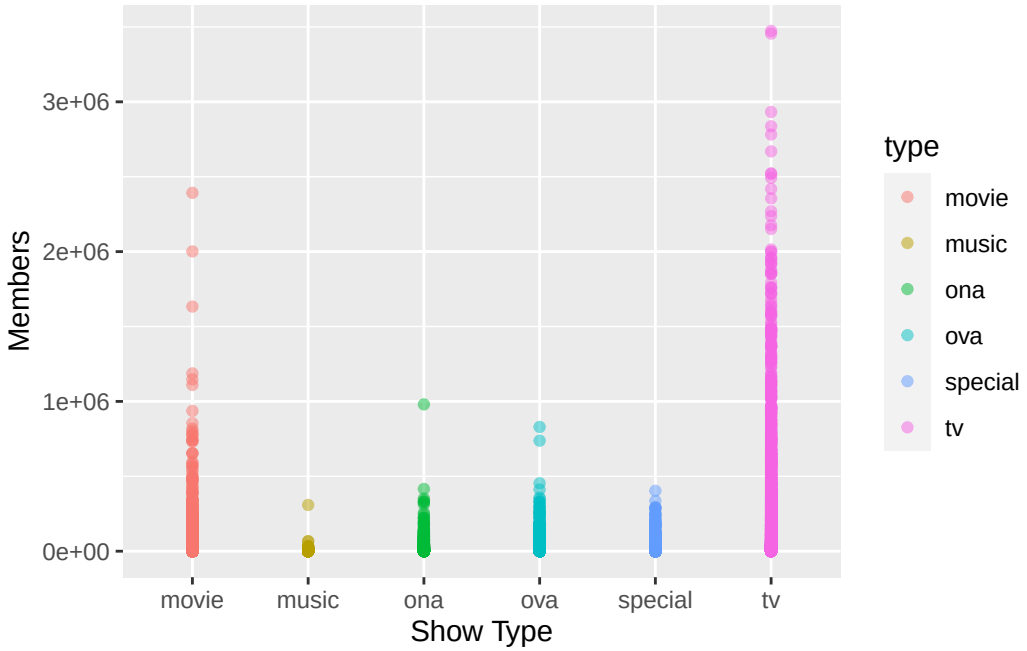

Figure 1: Visualization of how many MAL users viewed an anime based on the medium that it was presented in. The most popular mediums to produce anime in are movies and in TV shows, which is reflected with how many MAL users viewed anime in those specific categories. For reference, music anime refers to music videos presented in an anime artstyle, ONA is an original anime special included in DVD or Blu-Ray releases of an anime, OVA is an original video anime released exclusively online, and special anime are extra episodes attached to an already established series.

4.2 Episode Count

Episode count is also a factor that may affect how many MAL users would view a specific show. This is visualized in Figure 2, where the length of a series impacts MAL viewership. It is seen that many shows with episodes around 1, 12-13 and 24-26 have the most MAL viewership by users.

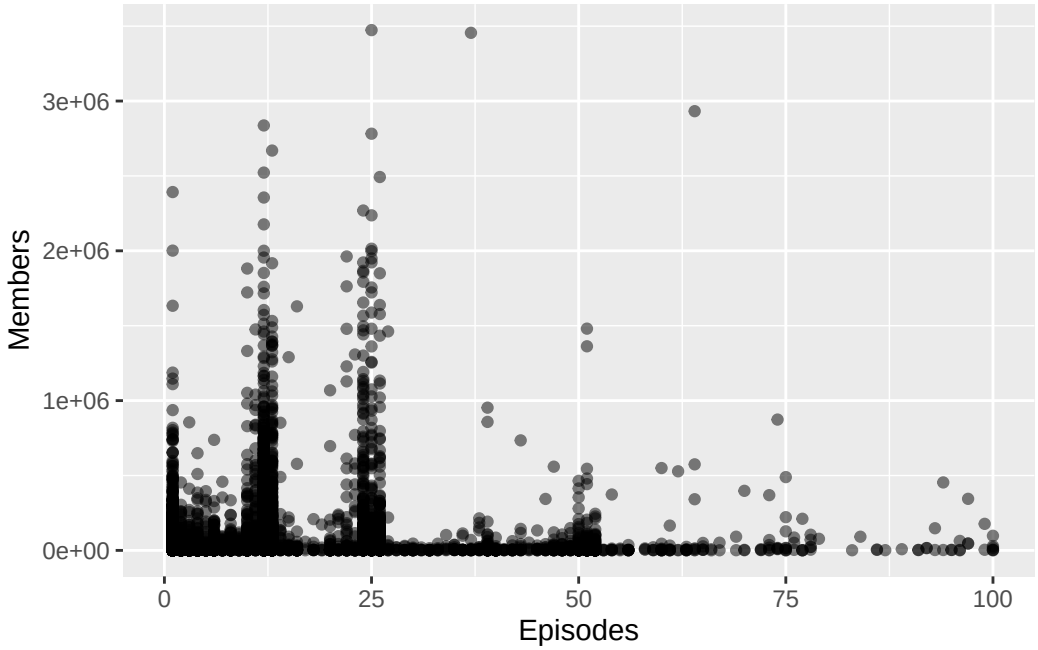

Figure 2: Visualization of how the episode count of an anime can influence MAL user viewership. There are areas with many specific peaks in viewership, most notably at the 1 episode, 12-13 episode, and 24-26 episode mark, which likely would pertain to movies and seasonal anime.

4.3 Source Material

The importance of the source material of an anime as explained in Section 3 is visualized in Figure 3. Many types of source material adaptations are present, with the most important ones to note being manga, light novels, and originals, which have the most MAL user viewership than any other source material disregarding outliers.

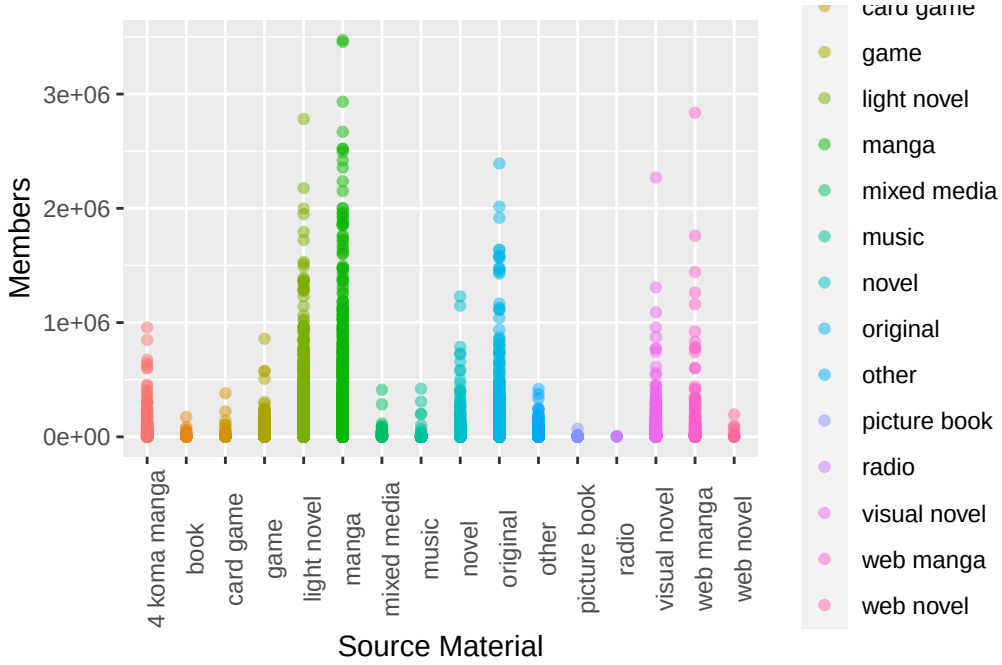

Figure 3: The influence of source material in MAL user viewership. Manga has the most pronounced and highest viewership of any other source material, followed closely by light novels and original anime.

4.4 Year of Release

A visualization of how viewership is impacted depending on how old an anime is can be seen in Figure 4. It can be seen that there are considerable more entries of anime in general, as more shows were produced in the 1970s. However, there does seem to be more MAL users that watch shows around past the year 2000 mark. Meanwhile, MAL viewership for anime produced before 1975 are rarely seen in comparison.

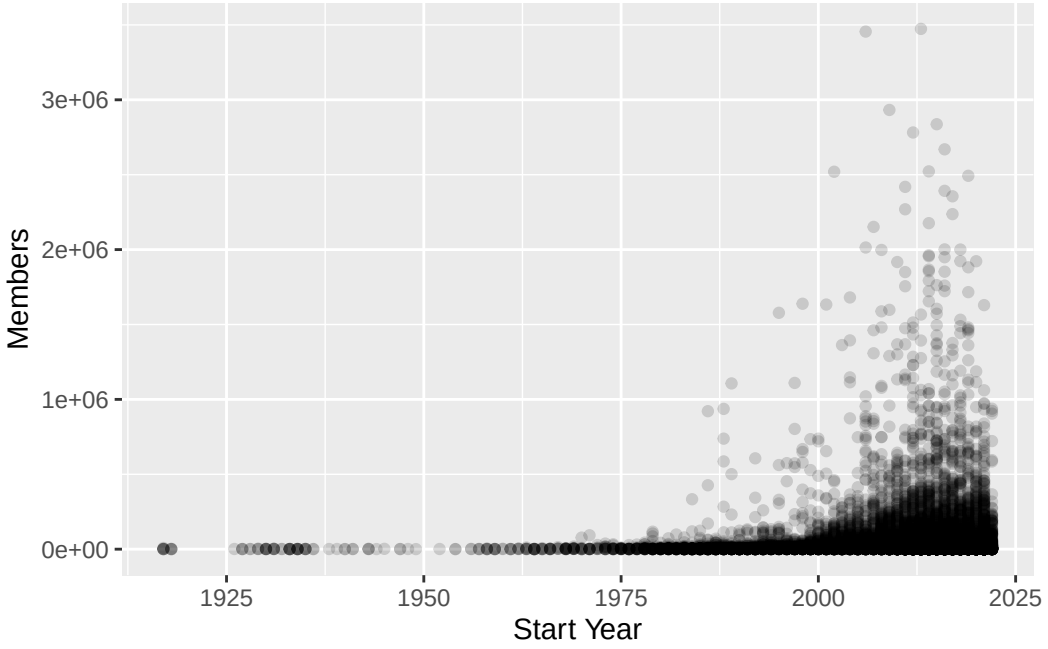

Figure 4: Visualization of the amount of viewers of a specific anime depending on the year of the anime's release. It can be seen that the amount of viewers of an anime increase considerably past the year 2000, but there are still many anime, especially between 1975 and 2000 with much less viewers.

5 Discussion

5.1 Show Type

5.2

5.3 Limitations and next steps

There are many limitations that are present in this statistical analysis. First, the depth of information is vast as there are many other variables that can be considered when determining the popularity of an anime. Factors from the production side such as the studios that produced the anime, to licensors that release the show on platforms to be watched on are also important for gathering attention to shows. In addition, factors that may give more information about what the show might be about such as the genres and themes may also play an important role as personal appeaal to specific genres may help gravitate MAL users to watch particular shows.

Another consideration that has to be made is in regards to

Appendix

A Datasheet for the Data Set

Extract of questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - This dataset was created to be used to aggregate the information of anime from the MAL database to be specifically used for statistical analysis. It can be hard to access publicly available information directly from the MAL and Jikan APIs, so this solution makes it easier for anyone to simply download and pursue exploratory data analysis with the plethora of information the data set provides.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by Andreu Vall Hernandez, a user on Kaggle. There was no specific intent in mind for the creation of this dataset but to make the information from the MAL and Jikan APIs publicly available. The dataset was not created for any specific entity.
    -

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - No organization funded the creation of the dataset.
    -

4. *Any other comments?*

    - None.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance in the dataset represent an anime entry within the MAL database since July 2022.

2. *How many instances are there in total (of each type, if appropriate)?*

- There are a total of 11,890 observations, each representing a different show.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset contains a sample of instances from the total shows available on the MAL database which can be accessed from the official MAL API. The last update to this dataset that was obtained was on July 2022, and is not indicative of all of the instances that may have been added between then and the time of this paper being published. Additionally, the dataset was cleaned to include instances that had all of the necessary variables to conduct statistical analysis on in this project.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance of data consists of the title of the show, score of the show, how many MAL users added it to their list, how many MAL users rated it, medium of the show, episode count, source material, start year, and studio that produced the show.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - The label of each instance is an anime.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - N/A

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Some relationships between individual instances are made explicit, for example when a show has a movie with the prefix of the movie title being the show's name. It is also made explicit for sequels of anime or spin-offs that include the main title of the original showl.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - N/A

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Some instances of the dataset do not contain any information, which affect how many shows could be analyzed in this report.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - All of the data within the dataset is publicly available information on the MAL website.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The dataset does contain keywords that may be offensive or insulting depending on the titles, themes, and descriptions of the instances.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - N/A

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - N/A

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset does not contain data that could be considered sensitive.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was collected via Kaggle and was directly observable, with variables and values explicitly visible.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was manually curated by combining two APIs, MAL and Jikan to obtian all of the relevant information within the dataset.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - N/A

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Only the original author, Andreu Vall Hernandez, was involved in the data collection process.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was last collected on July 25th, 2022, with older versions extending as far back as June 21st, 2022.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No ethical review processes were conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected from a third-party source, Kaggle.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - The individual who created the dataset was not notified about the data collection.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - The individual who created the dataset did not give any consent to the collection of the data that was gathered, however it was made available on Kaggle for use by the public, and included their GitHub repo to demonstrate how they collected the data for reproducibility.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - N/A

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - N/A.

12. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Variable names were renamed to be more clear, and the dataset was trimmed down to only include pertinent variables required for the analysis.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The raw data is accessible via the GitHub repo on the first page of the report.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Statistical programming language R (R Core Team 2023) was used to preprocess, clean, and label the data.

4. *Any other comments?*

    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - It is unknown if the dataset has been used by others as the dataset was made publicly available for anyone to use.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - The repository for how the dataset was created are available at https://github.com/andreu-vall/myanimelist-api-scraping

3. *What (other) tasks could the dataset be used for?*

    - The dataset can be used for analysis of different factors, or for different analytic reasons than the one described in this report. Much of the data pertains to information only present on MAL, so any kind of analysis on MAL could be done with this dataset.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - None.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- None, all of the data present within the dataset is publicly available information on the MAL webite, the dataset simply aggregates all of that information into one file.

6. *Any other comments?*

    - None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, as the GitHub repo will be made public.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed on the GitHub repo: https://github.com/justabando/malanalysis. There is no DOI.

3. *When will the dataset be distributed?*

- The datasets used in this paper will be distributed in April 2023.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset that was created for this paper is distributed under a MIT license. The original dataset is licensed under CC 1.0 Universal Public Domain Dedication.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No restrictions are placed on any of the datasets used or created for this report.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - None.

7. *Any other comments?*

    - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Justin Abando will host, maintain, and support the dataset at https://github.com/justabando/malana
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - The original creator of the dataset, Andreu Vall Hernandez, can be contacted at andreu.vallhernandez@gmail.com. The author of this paper, Justin Abando, can be contacted at justin.abando2@gmail.com.
3. *Is there an erratum? If so, please provide a link or other access point.*
   - N/A
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - Any updates will be posted on the GitHub repo used to host this paper which include the datasets.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - N/A.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Older versions of the datasets will be available on the GitHub repo.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - Anyone that wants to contribute to the original dataset should contact Andreu Vall Hernandez at andreu.vallhernandez@gmail.com.
8. *Any other comments?*
   - None.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.