

The multifaceted nature that determine an anime show's popularity involves (findings of the paper, WIP).*

Justin Abando

20 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 to determine why specific anime stand out, with some shows being rated lower in terms of quality, yet have more popularity based on MAL user responses. Data was retrieved to determine what factors could be responsible for a show's popularity. It was found that (results of analysis, WIP). 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 also useful at the community level. The rating and reviews that are given from individuals

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

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 its 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.

(EDIT) 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 results of paper, WIP).

(EDIT) 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 due to subject matter that may be difficult to process, whether

it would be highly philosophical, graphic, or leans in any other kind of extreme direction that may lose the interest of a casual viewer. 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. As such, the (factors that influence popularity that will be discussed on, WIP).

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.

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.
- **studios**: the names of the studios that produced the anime.

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. 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	Studios
1	Fullmetal Alchemist: Brotherhood	tv	64	manga	2009	Bones
2	Hunter x Hunter (2011)	tv	148	manga	2011	Madhouse
3	Shingeki no Kyojin Season 3 Part 2	tv	10	manga	2019	Wit Studio
4	Steins;Gate	tv	24	visual novel	2011	White Fox
5	Koe no Katachi	movie	1	manga	2016	Kyoto Animation
6	Kimi no Na wa.	movie	1	original	2016	CoMix Wave Films
7	Code Geass: Hangyaku no Lelouch R2	tv	25	original	2008	Sunrise
8	Shingeki no Kyojin: The Final Season	tv	16	manga	2021	MAPPA
9	Sen to Chihiro no Kamikakushi	movie	1	original	2001	Studio Ghibli

The main variables of interest as mentioned previously are medium of the show, episode count, source material, initial release year, studios, and licensors. These are all important and can possibly influence the popularity of an anime without giving any actual information of what the anime is about. First, a visualization of how viewership is impacted depending on how old an anime is can be seen in Figure 1.

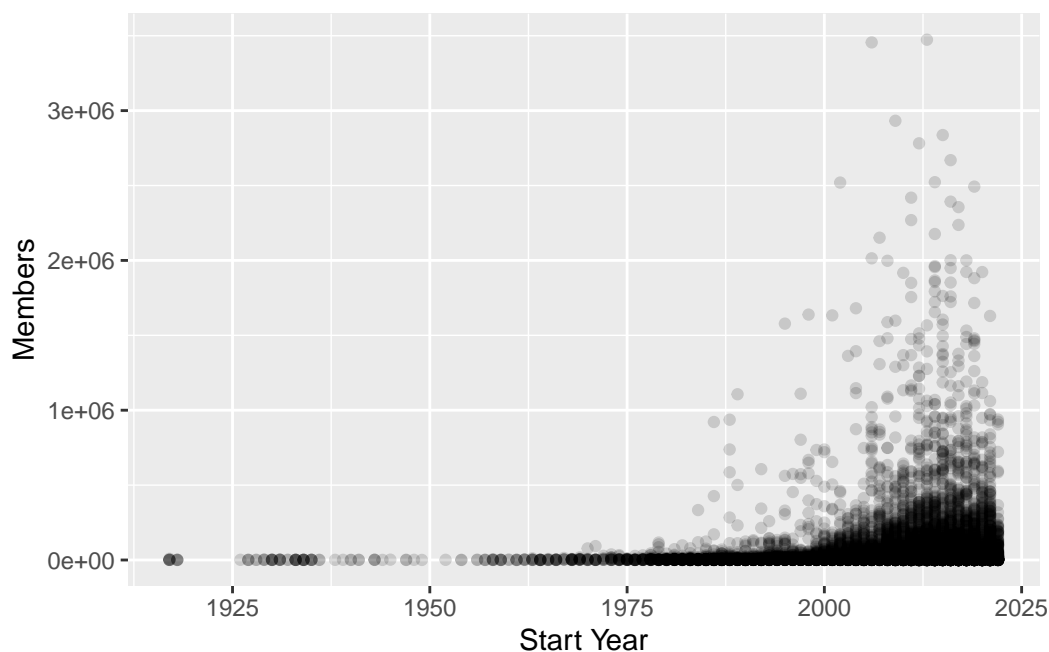


Figure 1: Visualization of the amount of viewers of a specific anime depending on the year of the anime's release. It can be seen that the amount of viewers of an anime increase considerably past the year 2000, but there are still many anime, especially between 1975 and 2000 with much less viewers.

It can be seen that there are considerable more entries of anime in general, as more shows were produced in the 1970s Figure 1. 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.

3 Model

A multiple linear regression model can be used to predict how each of the qualitative independent factors (show type, episode count, start year, source material, studios) can affect how

many users of MAL watch a given anime.

4 Results

5 Discussion

5.1 Limitations and next steps

Appendix

A Datasheet for the Data Set

Extract of questions from Gebreu 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.
2. *How many instances are there in total (of each type, if appropriate)?*
 - There are a total of 24,165 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 is a sample of instances. It is obtained through convenience sampling from the larger set of Black Desert Online players on the North American and European servers. Since the dataset is obtained through convenience sampling, it is not representative of the total population in terms of geographical coverage or any stratification. The choice of convenience sampling is due to the lack of APIs and useful tracking tools provided by Pearl Abyss, the developers behind Black Desert Online.
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 a player’s age, gender, total playtime in the game and on their main character, their rated importance of various subscription items in the game, and their total spending on various types of items in the game.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - the gender is either male, female, or other.
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.*
 - There is lack of female and other genders in certain age groups in the dataset due to insufficient sample size.
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.*
 - There are no relationships between individual instances.
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.*
 - None.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- The sample size is unequal for female and male respondents. This unequal sample size does negatively affect the assumption of equal variances. Furthermore, lack of detailed statistic tracking tools and APIs provided by Pearl Abyss—the developer of Black Desert Online—means that playtime and most monetary questions in the survey rely on the respondents’ best estimates.
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.*
 - Breakdown of the monetary transactions of each player might be considered a confidential data by Pearl Abyss. However, the survey was designed to be as transparent with the respondents as possible, and no data that could be linked back to the respondents were included in the survey nor in the final paper.
 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 not contain any offensive or insulting data.
 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.*
 - The datasets are split by age groups and gender as self-identified by the respondents.
 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.*
 - It is entirely impossible to identify individuals from the dataset.
 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 datasets do not contain data that might be considered sensitive in any way.

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 through a survey distributed to the subreddit page of Black Desert Online, the official website's forum of Black Desert Online, and the steam discussion page of Black Desert Online. The data was reported by the respondents based on their best estimates.

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 collected through a survey distributed to the subreddit page of Black Desert Online, the official website's forum of Black Desert Online, and the steam discussion page of Black Desert Online. The data was reported by the respondents based on their best estimates. Google Forms is used to produce the survey.

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

- Convenience sampling was used due to lack of detailed statistic tracking tools and APIs provided by Pearl Abyss.

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)?*

- No one but the author is 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 collected over the span of four days from April 18, 2022 to April 21, 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. Subreddit moderators were consulted for permission of posting the survey.
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 through a survey distributed to the subreddit page of Black Desert Online, the official website's forum of Black Desert Online, and the steam discussion page of Black Desert Online.
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.*
 - Respondents were informed of the purpose of the data collection: "I wish to study the different monetization methods that BDO employs. Due to the lack of proper APIs provided by Pearl Abyss, I have resorted to surveys instead."
 - "To be clear, there will not be any information present in the final paper that can be used to link the data back to the original respondents."
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.*
 - Since data collection is through a survey, consent was explicitly provided when respondents clicked the survey link.
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).*
 - Respondents can revoke consent by simply exiting the survey.
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.*
 - Certain variables names were shortened in Google Spreadsheet before the dataset is used for analysis in this paper.
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 in the github link provided in this paper.
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 (**citeR?**) was used to preprocess the datasets.
 - Google Spreadsheet is publicly available for use by registering a Google account.
4. *Any other comments?*
 - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - At the time of writing this, only this paper has utilized the dataset.
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 this paper, which uses the datasets, are available at https://github.com/zhan7818/bdo_monetization.
3. *What (other) tasks could the dataset be used for?*
 - The datasets could be used to modeling or finding relationship between total in-game spending and age groups.
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.*
- The datasets should not be used in an effort to harass or discriminate against the general Black Desert Online community nor should it be used to misrepresent information about the said community.
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.*
- Since there dataset is publicly available through the github link, yes.
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 Tian Yi Zhang’s github page, along with this paper: https://github.com/zhan7818/bdo_monetization.
3. *When will the dataset be distributed?*
- The datasets used in this paper is distributed in 2022.
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 is distributed under a MIT license.
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 the dataset.
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?*

- Tian Yi Zhang is hosting the dataset on the github link at https://github.com/zhan7818/bdo_monetiz

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- The creator of the datasets, Tian Yi Zhang, can be contacted at tianyi.zhang6546@gmail.com.

3. *Is there an erratum? If so, please provide a link or other access point.*

- As of writing this datasheet, no explicit erratum have been spotted.

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 page 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 (if there will be any) will be kept as a record on github.

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.*

- One should contact Tian Yi Zhang, the author of the dataset.

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/>.
- 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 Golemund, 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*.