Anime Data Analysis and Recommendation System

Group 15

Chinmaya Singal	180207	chinmaya@iitk.ac.in
Dipesh Khandelwal	180249	dipeshk@iitk.ac.in
Rythm Agarwal	180636	rythm@iitk.ac.in
Sakshi	180653	sakshisa@iitk.ac.in
Sarthak Dubey	180674	srthkdb@iitk.ac.in

Contents

1 Introduction

Anime refers to hand-drawn and computer-animated animation originating primarily from Japan. It also includes other animation such as Avatar: The Last Airbender and RWBY which are inspired from the traditional Japanese anime but produced outside Japan. Anime as an entertainment media has been amassing immense popularity and gaining more audience in the recent times.

In this project, we have analysed the data available on MyAnimeList and tried to get some insights related to the growth of anime over time, genres, studios, etc. Along with that, we have also built anime recommendation systems using different models to provide both content based and collaborative filtering based recommendations to users.

2 Data Retrieval

The dataset we used for our project is the MyAnimeList dataset.
To retrieve the data corresponding to each anime and users on MyAnimeList, we used the official MyAnimeList API, the documentation for it can be found at MAL API. The process of data retrieval and the problems we faced have been explained below:

- **Anime data:** The first problem we noticed while trying to retrieve the data for every anime was that the api requires us to input an anime id but not every integer anime id (upto a certain limit) is valid which posed a problem in retrieving data. To counter this problem, we searched for a source which contains the valid anime ids which can be found here. We scraped the valid anime ids from the given link and used those ids to make API calls to the official MAL API and collect data for each anime listed on MAL in JSON format.

- **User data:** The data for a user containing a list of anime they have watched along with their rating of each anime can be easily fetched using the MAL API. But in this case we ran into a problem that we don't have a list of valid usernames. In order to retrieve a list of usernames, we used a Breadth First Search algorithm on the network of users and their friends starting from the account of one of our group members. Since the official MAL API doesn't currently provide support for fetching friend list of a user, we had to use an unofficial API whose documentation can be found at Jikan API. Once we had retrieved enough usernames, we fetched their anime lists using the official MAL API.

3 Recommendation Systems

3.1 TF-IDF based Recommendation System

Term Frequency — Inverse Document Frequency (TF-IDF) based recommendation systems are content based recommenders. This method primarily uses the synopsis of an anime to provide recommendation by assigning importance to words mentioned in synopsis with more importance to words mentioned less commonly among all synopsis.

- **Recommendation using Synopsis:** First we built a recommender which only used synopsis of an anime [1] to suggest similar anime. To do so, first we lemmatised each synopsis using tools abailable in the nltk library of python. Then we vectorised the synopsis using TfidfVectorizer from sklearn and found the cosine similarity between pairs of vectors. In order to make recommendations, we pick the anime having synopsis vector most similar to a given anime in terms of cosine similarity. By default, we output top 10 recommendations provided by the engine. The code for this can be found in tf_idf_synopsis_recommender.py.

  One problem we noticed with the recommendations provided is that the genre of recommended anime can be completely different from the given anime due to similar terms in synopsis. To counter this problem, we improved on the model as explained below.

- **Recommendation using Synopsis and Genre:** In addition to the synopsis based recommender, we also build a genre based recommender [5] using similar method but this time we tokenised it based on subsets of genres since the order of genre doesn't matter. The results of genre based recommender alone weren't promising so we combined the two scores of each anime obtained from both recommenders and provided recommendation of those anime which have high similarity score in both genre and synopsis. The code for this can be found in tf_idf_syn_genre_recommender.py

  This gave us better results than before but we noted that the recommendations generally had a low rating since similarity was found with obscure and not so good anime which motivated us to improve the model further as explained below.

- **Recommendation using Synopsis, Genre and Rating:** In order to incorporate rating to some extent into our recommender, we multiplied the obtained similarity score by the average rating of the anime being recommended. In addition to this, we also dropped any anime having a rating of less than 7 since that is the threshold below which generally anime are not very good (This threshold is based on domain knowledge). The code for this can be found in tf_idf_syn_gen_rate.py.

  This recommender provided us much better results than the previous two since it also takes into account the rating of anime to some extent. The importance of rating was kept comparatively low so that recommendations can be actually useful to users rather than just recommending very popular anime which they are likely to have already watched.

3.2 KNN Item-Based Collaborative Filtering

The collaborative filtering approach builds a model from a user's past behaviors (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in.

One of the main challenges in building a recommendation system is that many of the animes do not have enough rating which is common with anime with low popularity. Another challenge is many users are not active in providing ratings which makes it difficult to analyze user behavior.

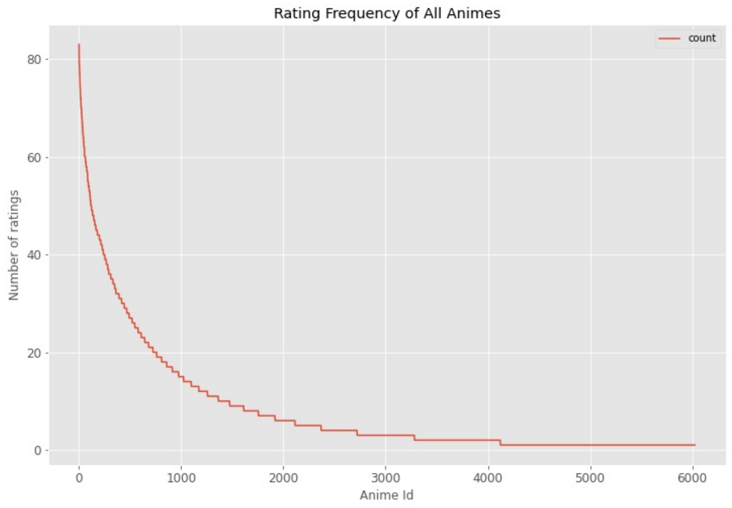

Figure 1: Rating Frequency is a “long tail” distribution. Number of ratings available for each anime sorted in descending order.

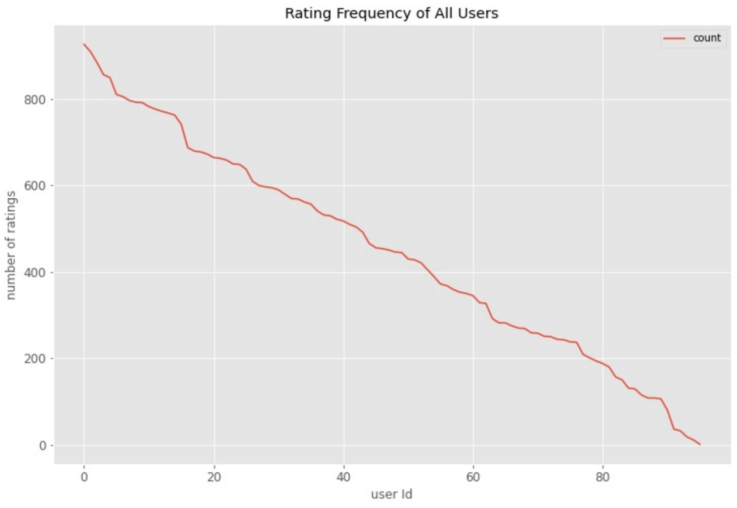

Figure 2: Number of ratings provided by users sorted in descending order.

For data filtering, we filter out animes that have very low popularity. On user data, we filter out inactive users which we judge by the number of animes they have rated. This step avoids noise in the result and also makes it easier to process the large dataset.

We used item-based collaborative filtering with KNN [3] to build our recommendation system. KNN is a non-parametric, lazy learning method. It uses a database in which the data points are separated into several clusters to make inferences for new samples. We used cosine similarity for the nearest neighbor search.

3.3 Alternating Least Square (ALS) Matrix Factorization based Collaborative Filtering

The Alternating Least Square (ALS) [6] [4] algorithm is a matrix factorization approach that runs in parallel. ALS is built for large-scale collaborative filtering tasks and is implemented in Apache Spark ML. ALS does a decent job of dealing with the Ratings-data's scalability and sparseness, and it's simple and scales well to very big datasets.

- **Shortcomings of KNN:** Because KNN-based recommendation systems suffer from popularity bias and item cold-start issues, we employ ALS. The recommender's popularity bias means that the animes with the most interactions are recommended, and the item cold-start problem means that new content with no or very few interactions is not recommended.

- **How does ALS deal with these flaws?** To tackle KNN's shortcomings, we built our recommender system using the Matrix Factorization technique. As the model learns to factorise rating matrices into user and anime representations, the popularity bias is eliminated, allowing the model to predict better personalised anime ratings for users. Also, the item cold-start problem is handled because, according to matrix factorization, lesser-known animes can have rich latent representations just like popular movies, which improves the ability of recommenders to recommend lesser-known animes.

- **Working of the ALS recommender** We used Python's PySpark library to implement a distributed recommender system based on ALS. Users rate certain animes, and this information is used as training data for the model. This data is then used to train the ALS model. Based on the user's ranking of animes, the programme generates the top N movie recommendations for the user.

The ALS model can produce unpopular recommendations, which is useful for recommending fresh information to users and keeping them engaged. We can improve the model even further by utilising hybrid models that mix KNN and ALS recommendations to produce recommendations for the user that include both popular and new content.

3.4 Item-Based Collabortive Filtering with Singular Value Decomposition (SVD)

We created an item-based collaborative filtering model using Singular Value Decomposition (SVD). Item based recommendation is based on the user-item rating (here, anime rating). The assumption in collaborative filtering is that people who have liked an item in the past will also like it in the future. This approach builds a model based on the past behaviour of users. The user behaviour is modeled by their ratings for various animes. The model is an association between the user and the items (anime) which is used to predict other items the user maybe interested in.

We have used Singular Value Decomposition as a collaborative filtering[2] approach in our anime recommender system. SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions. However, we are only concerned with the matrix factorization part keeping same dimensionality. The matrix factorization is done on the user-item ratings matrix.

We use cosine similarity function for determining the most similar animes and return the top N (as asked by the user) animes based on this similarity.

Issues with SVD-based Collaborative Filtering:

1. **Data Sparsity:** A huge problem caused by the data sparsity is the cold start problem. As collaborative filtering methods recommend items based on users' past preferences, new users will need to rate a sufficient number of items to enable the system to capture their preferences accurately, and thus provides reliable recommendations. Similarly, new items also have the same problem.

2. **Gray sheep:** Gray sheep refers to the users whose opinions do not consistently agree or disagree with any group of people, and thus do not benefit from collaborative filtering.

4 Data Analysis

4.1 Performance of Related Anime

Related anime refers to the anime which are related to a given anime. These may include different seasons of the given anime, OVAs, specials, movie, etc. We tried to infer whether the rating and popularity of related anime is dependent on the given anime. In general we would expect it to be the case since a popular or well acclaimed anime is likely to have a popular and well acclaimed second season.

To test our hypothesis, we used the Pearson correlation coefficient with the null hypothesis that the performance of related anime is not dependent on the performance of given anime. We obtained the following results:

- **Rating:** We tested whether the rating of an anime and the set of related anime have a correlation. The Pearson Correlation Coefficient turned out to be approximately 0.67 indicating a strong correlation between the rating of a given anime and the related anime as we would expect.

  The p-value of the test also turned out to be 0, hence we should reject the null hypothesis. Although, in this case, p-value is not very indicative since our dataset is very large so the significance interval is very small.

- **Popularity:** We tested whether the popularity of an anime and the set of related anime have a correlation. We used the number of users who have watched a given anime as the measure of popularity. The Pearson Correlation Coefficient turned out to be approximately 0.19 indicating a weak correlation between the popularity of a given anime and the related anime. This indicates not as many people watch related anime, possibly because people who didn't like the first season are less likely to watch any subsequent related releases.

The p-value of the test turned out to be 0, hence we should reject the null hypothesis. Although, in this case, p-value is not very indicative since our dataset is very large so the significance interval is very small.

4.2 Rating Trends based on starting date

On careful analysis and comparison of both the graphs in Figure 1 and Figure 2, we can observe that the anime ratings saw a decreasing trend at the starting and had their worst ratings around 1945. After 1945 the ratings have shown increasing trends which go on rising continuously. Also, most of the anime produced are of type movie, tv or ona and this can be observed through the graph too. The rating trends of all the anime follows almost the same trajectory as the anime trends in which only movie, tv and ona type anime are included.

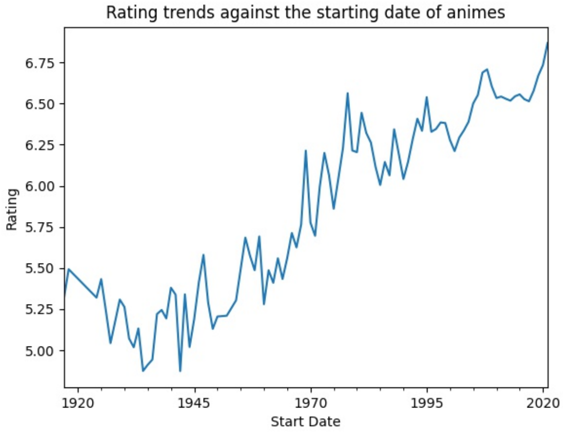

Figure 3: Rating trends against starting date of animes

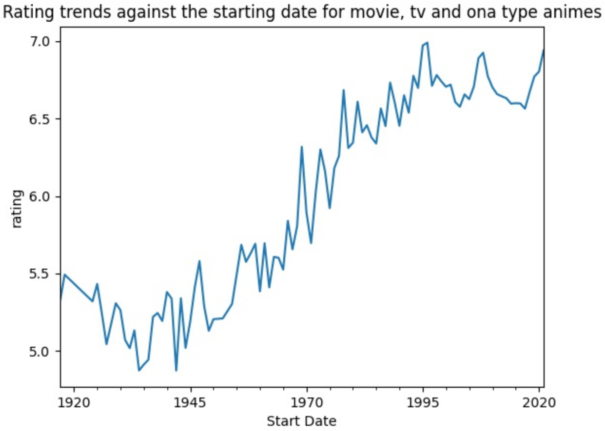

Figure 4: Rating trends against starting date of movie, tv and ona type animes

4.3 Trends based on Media Type

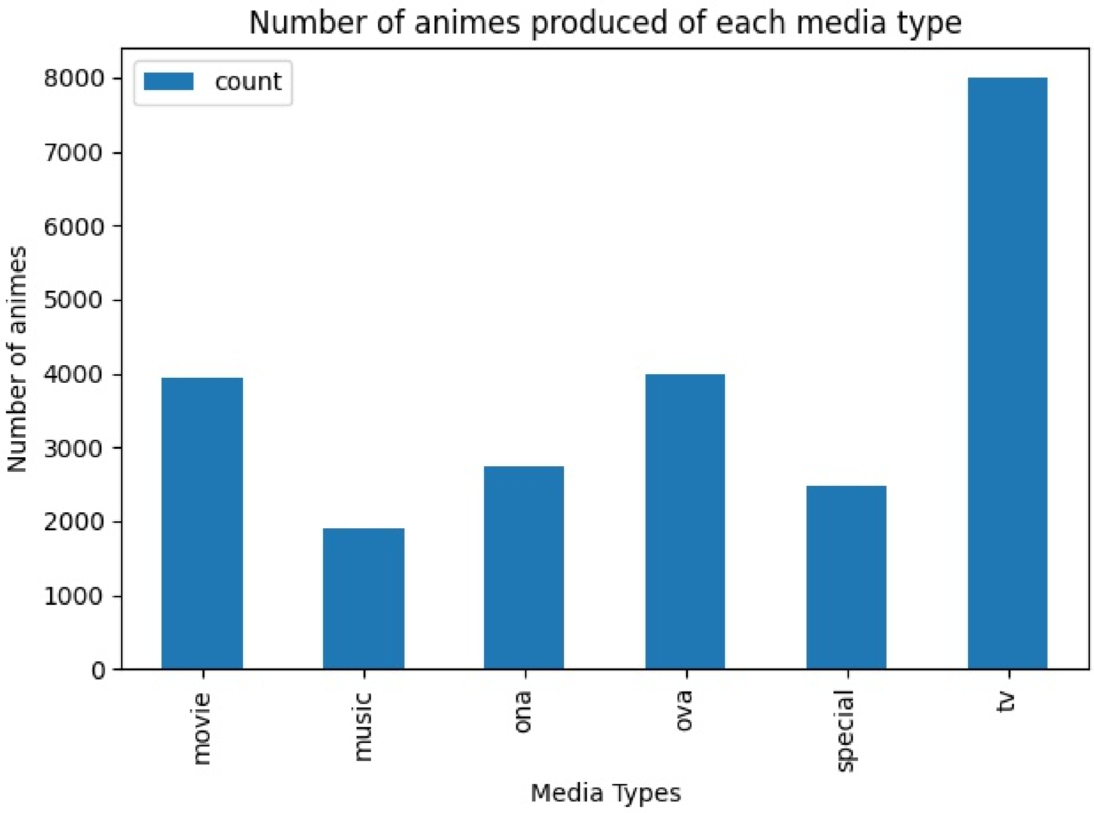

Figure 5: Number of animes produced of each media type

Insights:

1. Clearly, number of animes produced for TV are much more than animes produced of any other media types. Number of OVA type animes, which is the second most produced type is almost half of the number of the number of animes produced for TV.

2. Music type animes are the least produced animes.

3. Media types in the decreasing order of number of animes produced is:

$$\text{TV} > \text{OVA} > \text{Movie} > \text{ONA} > \text{Special} > \text{Music}$$

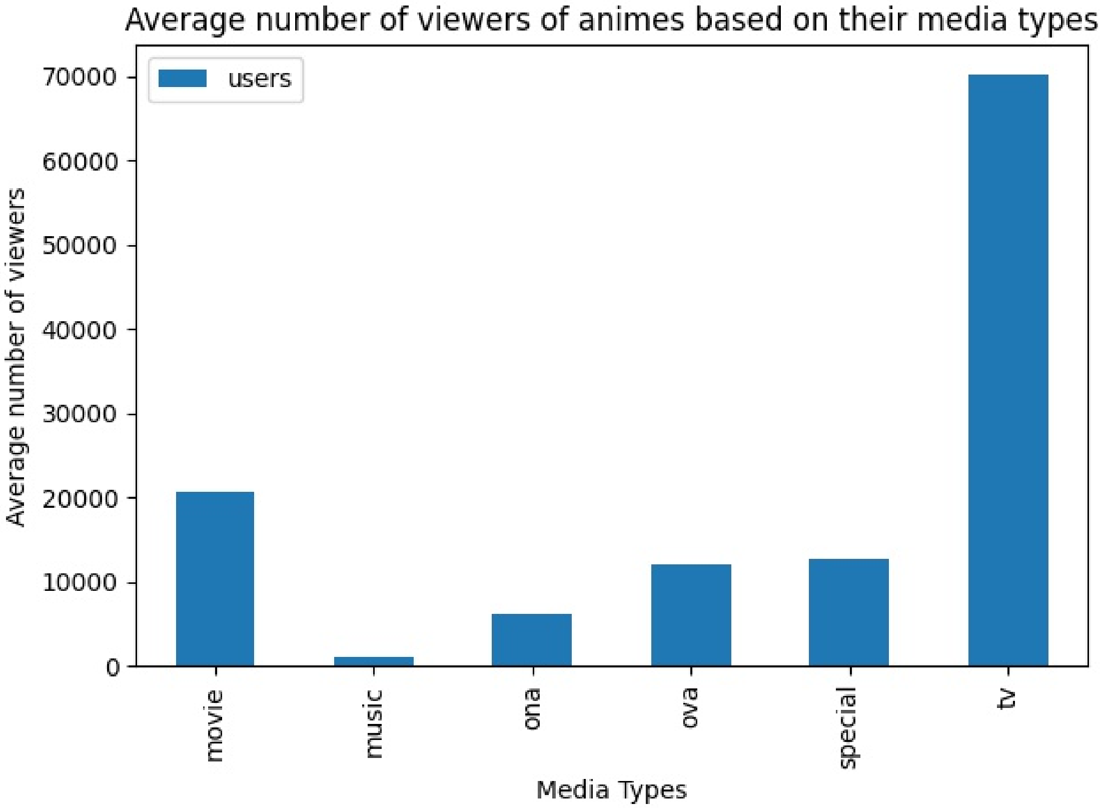

Figure 6: Distribution of number of viewers of animes based on their media types

Insights:

1. The average number of viewers of TV-type animes are far greater than any other types. This also explains the observation from the above graph that the number of animes produced of TV-type are much greater than any other types as TV-type animes attract most viewership.

2. The second most number of average viewers are attracted by Movie type animes, although they are significantly less than the average number of viewers of TV-type animes.

3. Media types in the decreasing order of average number of viewers is:

$$\text{TV} > \text{Movie} > \text{Special} > \text{OVA} > \text{ONA} > \text{Music}$$

4.4 General Trends

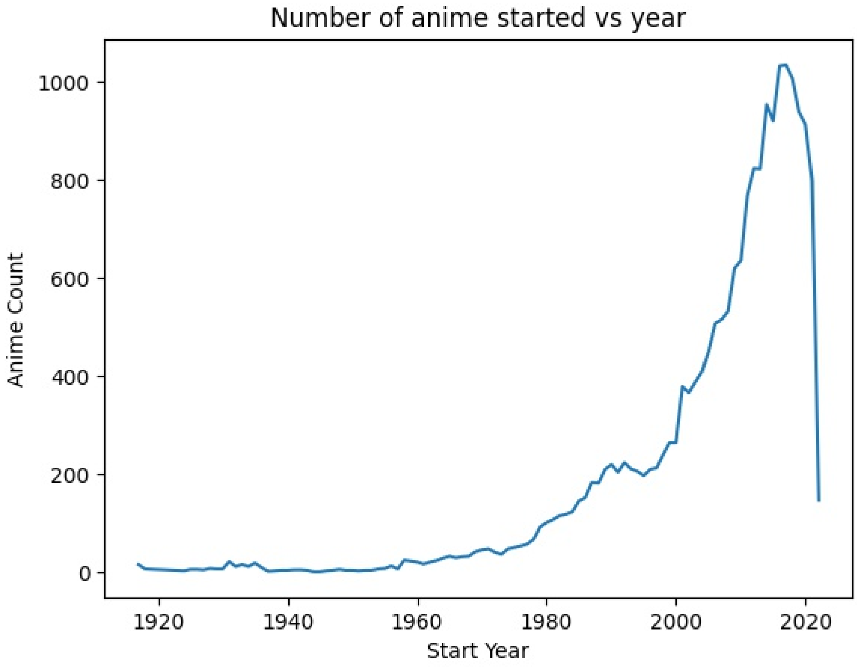

Figure 7: Temporal analysis of the number of animes released each year

Insights:

1. The number of animes started per year has been increasing steadily from almost no new animes published during 1920 to more than a thousand new animes started in 2019. The rate of increase was almost zero from 1920 to 1960 since those were the early years when anime initially started getting produced.
2. The rate of growth then boosted around the $1960s$ and then again during the $2000s$.
3. The dip in the number of animes started during $2020-21$ is due to the $COVID-19$ pandemic.
4. The sharp drop at the end is because of the upcoming announcements for animes to be started in 2022.

4.4.1 Rank vs. Popularity

We analyzed the field rank and popularity of the anime data to see if they are related. We calculated the correlation of the two fields using the Pearson method, a statistical measure that expresses the extent to which two variables are linearly related. We observed that correlation $= 0.9104$. This value indicates that the two fields are highly correlated.

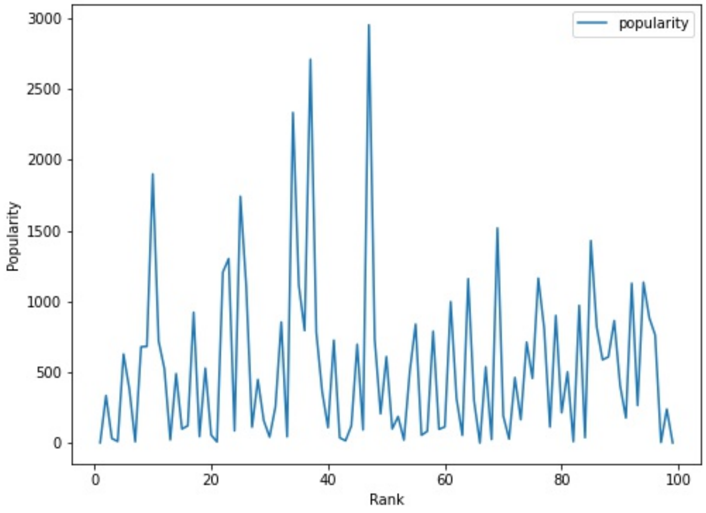

Figure 8: Plot depicting the trend of rank vs. popularity for top 100 ranked anime.

Mean of difference of $|rank - popularity| = 499.34$ for top 100 ranked anime.
Standard deviation of difference of $|rank - popularity| = 571.57$ for top 100 ranked anime.

We also plotted rank vs. popularity for all the anime data we have. Due to high variance in the data points, to better understand the trend, we smoothed the curve by plotting the average popularity in intervals of 200 (Total number of data points = 17314).
Mean of difference of $|rank - popularity| = 2465.72$
Standard deviation of difference of $|rank - popularity| = 1858.63$

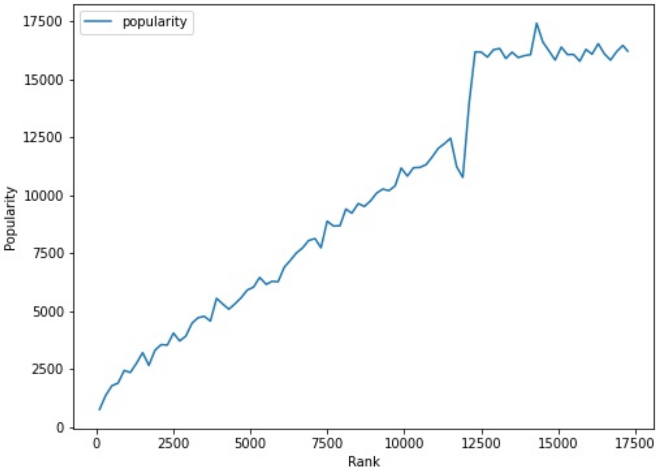

Figure 9: Plot depicting the trend of rank vs. popularity for all animes.

4.5 Studio Trends

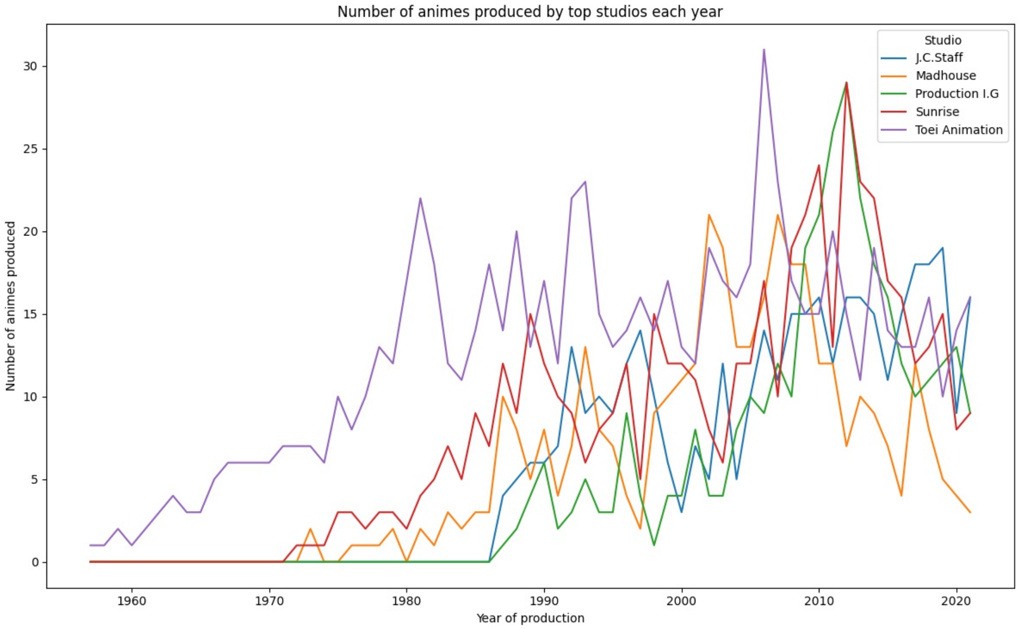

Figure 10: Trend of number of animes produced by top 5 studios every year where top 5 studios are choosen in terms of number of animes produced

Insights:

1. Toei Animation, Production I.G, Sunrise have produced more than 25 animes a year atleast once till now.
2. Toei Animation is one of the oldest anime producing studio and J.C. Staff and Production I.G are the latest ones among the top-5 studios in terms of number of animes produced
3. Madhouse has seen a much sharper decrease in the number of anime produced as compared to other studios in this list in the past decade.

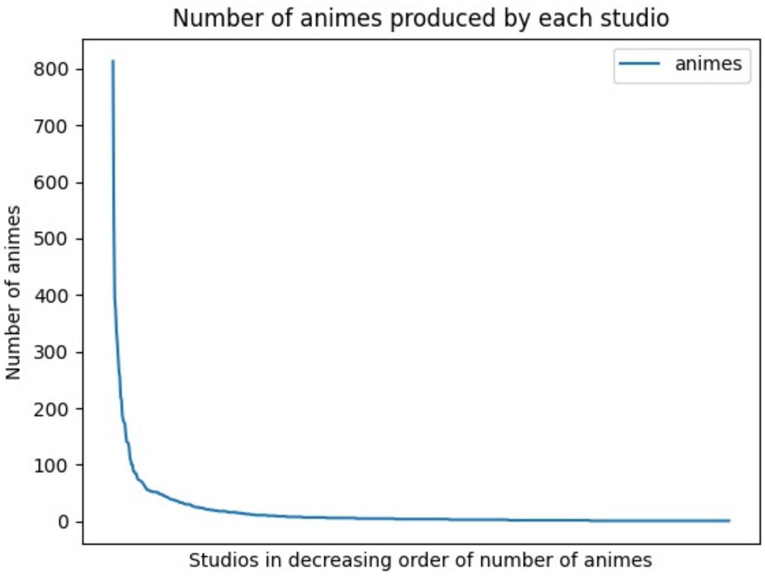

Figure 11: Number of animes produced by each studio

Insight: The number of animes produced by each studio follows power-law distribution which is commonly seen in many social settings.

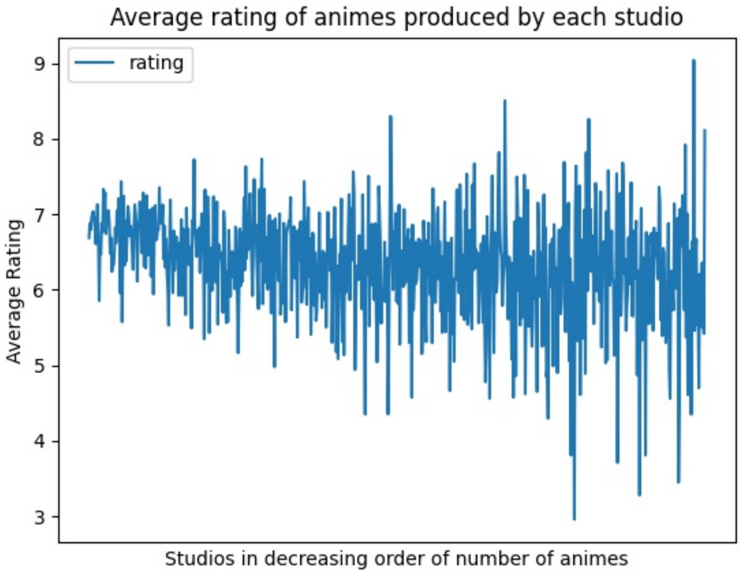

Figure 12: Average rating of animes produced by each studio

Insights: Average rating of animes produced by top studios is less variable than that of studios which have produced very less number of animes. This can be easily noticed from the graph as the graph is more zig-zagged with large differences between extremes towards the right side in comparison to studios on the left side. The reason for this is likely that the productions by smaller studios are either a hit or a miss.

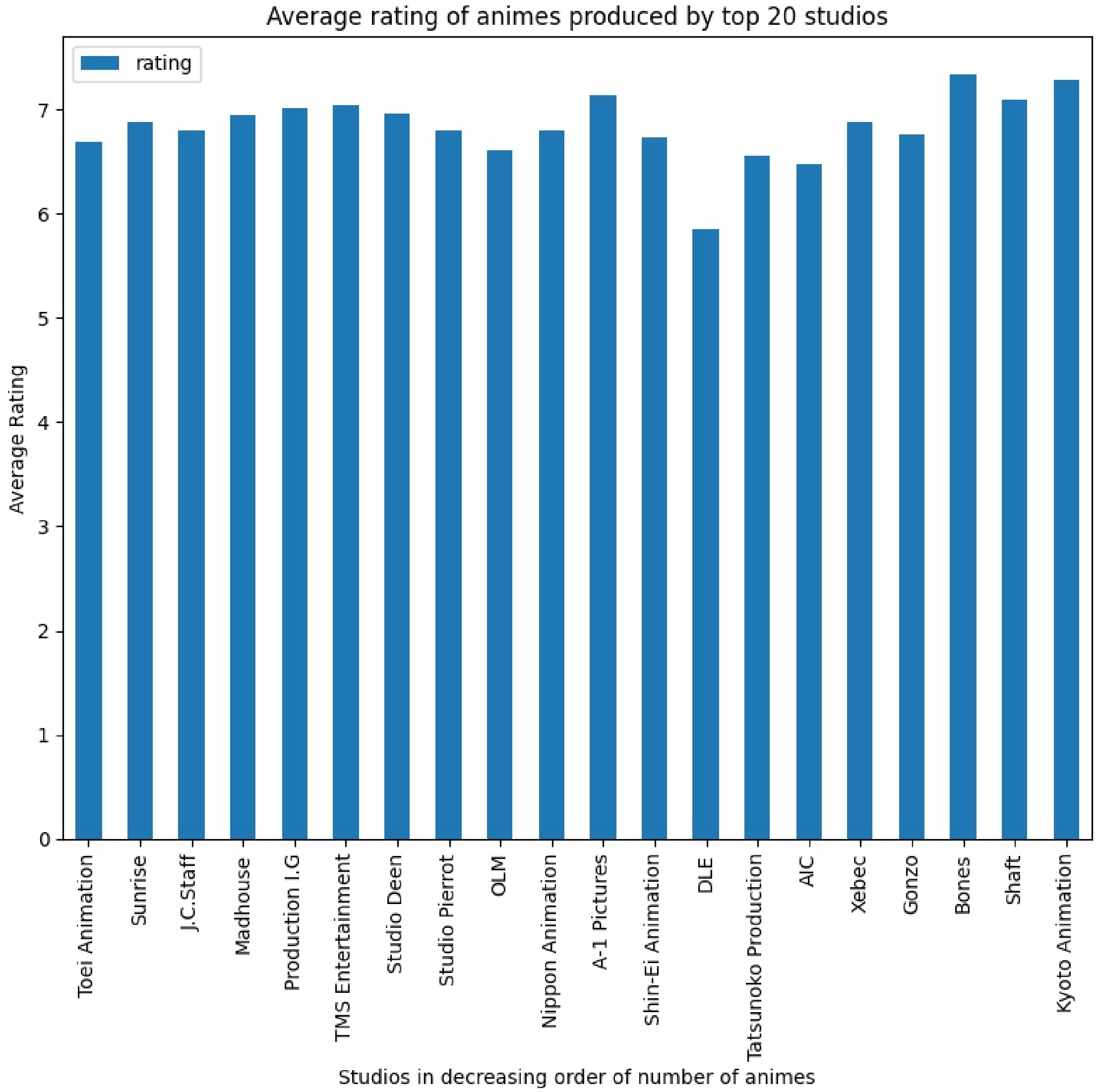

Figure 13: Average rating of animes produced by top 20 studios where top 20 studios are choosen in terms of number of animes produced

Insights:

1. DLE is the lowest rated studio among the top-20 studios in terms of number of animes produced. Average rating of animes produced by this studio is less than 6, while for other studios it lies well above 6 (mostly in the range of 6.5 to 7).

2. Bones is the highest rated studio among the top-20 studios in terms of number of animes produced.

3. The average rating of animes produced by top-20 studios is very close to each other and a significant difference cannot be observed between them.

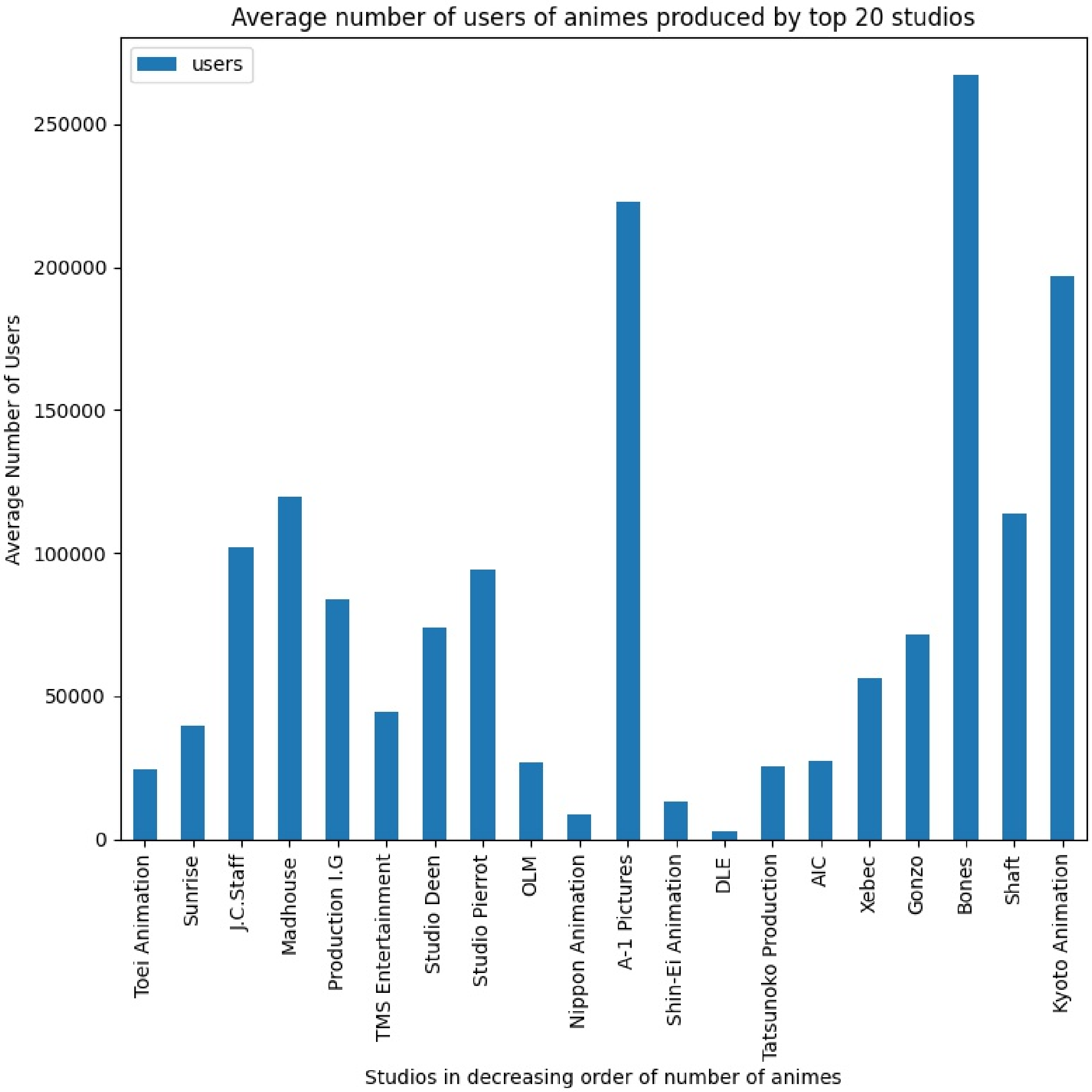

Figure 14: Average number of users of animes produced by top 20 studios where top 20 studios are choosen in terms of number of animes produced

Insights:

1. The average number of users of studios do not follow the trend of number of animes produced by the studio.

2. Clearly, Toei Animation which has produced more than 800 animes have very less average number of users in comparison to Bones which comes at 18th position in terms of number of animes produced but has the highest average number of users in this list.

3. Anime produced by DLE and Nippon Animation have very less viewership despite the fact that they are among top-20 studios in terms of number of animes produced.

4.6 Age Rating Trends

Based on its content, an anime can be rated as either $g, pg, pg13, r, r+$ or rx. We analyzed the trends of change of age rating on several variables and generated the following plots:

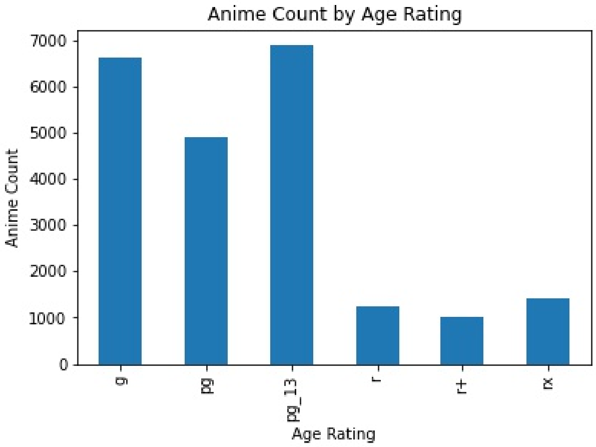

Figure 15: Number of anime against age rating

Insights:

1. The number of child-friendly animes released is more than five times the number of animes released for adults.
2. The most number of animes are $pg13$ rated while the least are in $r+$ rated.

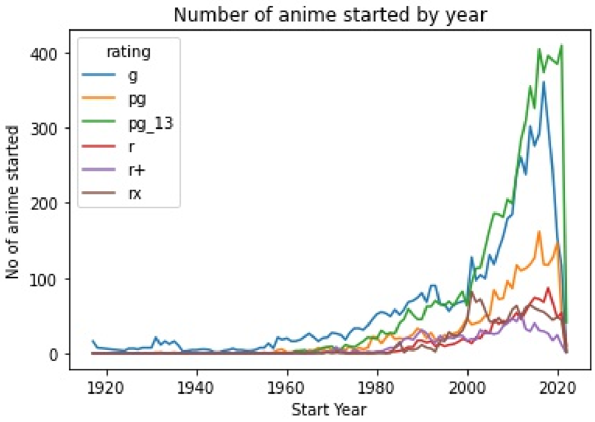

Figure 16: Temporal trends of number of anime started against start year for each age rating

Insights:

1. Growth in the number of animes published is the most for *pg*13 and *g* rated animes, moderate for *pg* rated animes, and the least for *r*, *rx*, and *r*+ rated animes.

2. It is evident from the curve that the growth in animes produced for adults is low when compared to anime for children.

3. Again, the sharp drop at the end of the graph is due to future announcements of animes to be released in 2022.

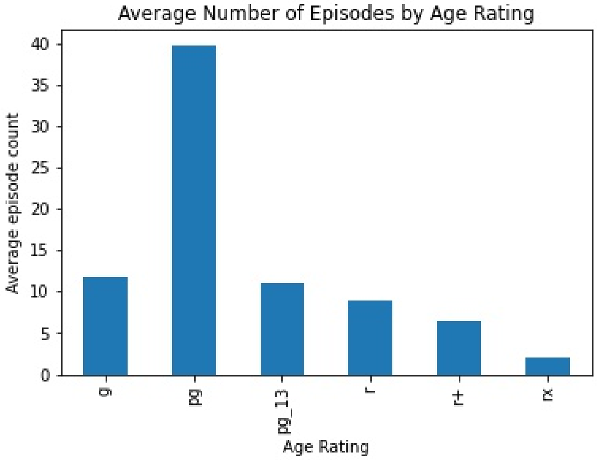

Figure 17: Average episode count against age rating

Insights:

1. The average number of episodes is the highest for *pg* rated animes, more than three times of any other age rating.

2. The average number of episodes decreases steadily from *pg*13 to *rx* rating.

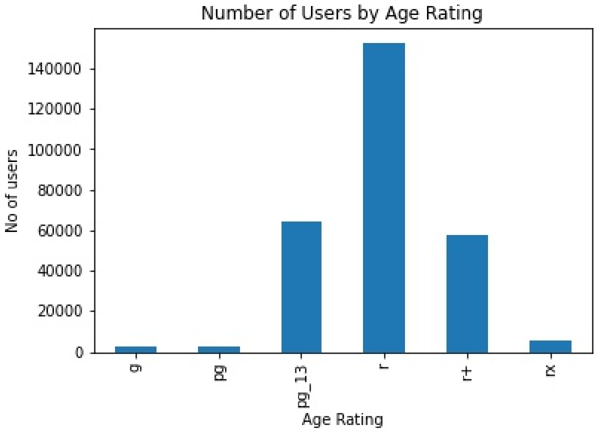

Figure 18: Number of users against age rating

Insights:

1. The above figure indicates that the maximum number of users add r rated animes in their list, and very few users add animes rated g, pg, and rx in their list.

2. This does not necessarily indicate the viewership of each age rating because most of the children likely do not use *MyAnimeList*, and hence the less number of users who watch g and pg rated anime.

4.7 Source Trends

We analysed whether the source material of the anime has any relation to the performance of the anime. To that extent, we plotted the average rating, average number of users (indicative of popularity) and the number of anime produced over time for each source and noted the following:

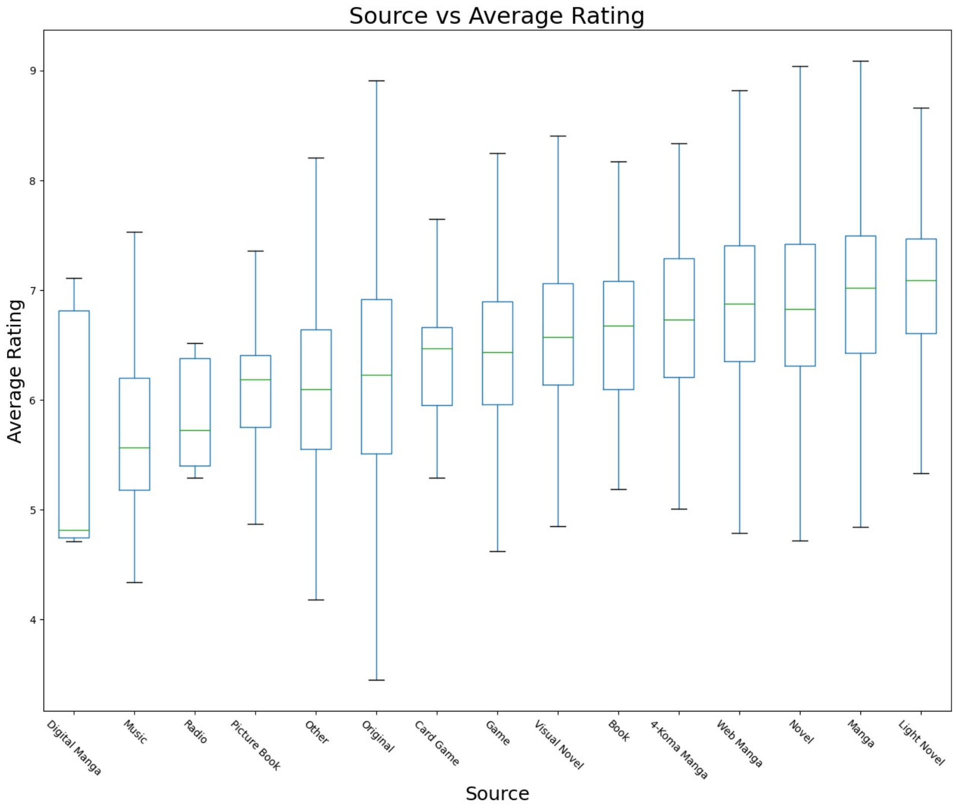

Figure 19: Average rating of anime corresponding to the source material

We note that the anime having Light Novel or Manga as their source material have the highest average rating whereas digital manga has the lowest. This is also expected since Manga is the most common source for anime besides original anime and anime sources from light novels have also seen a rise in the recent decade thanks to their performance.

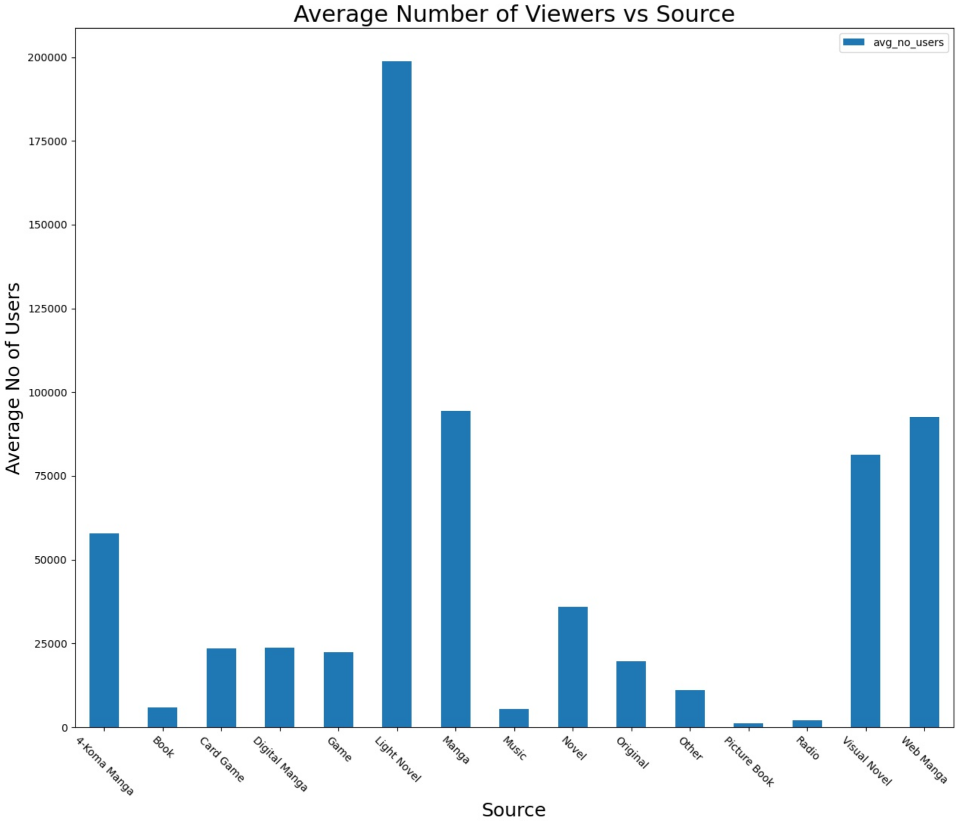

Figure 20:

Here, we note that the anime sourced from light novels enjoy a much higher average viewership than any other media. Anime sourced from Manga, Web Manga, Visual Novel and 4-Koma Manga also have fairly high viewership as compared to other sources.

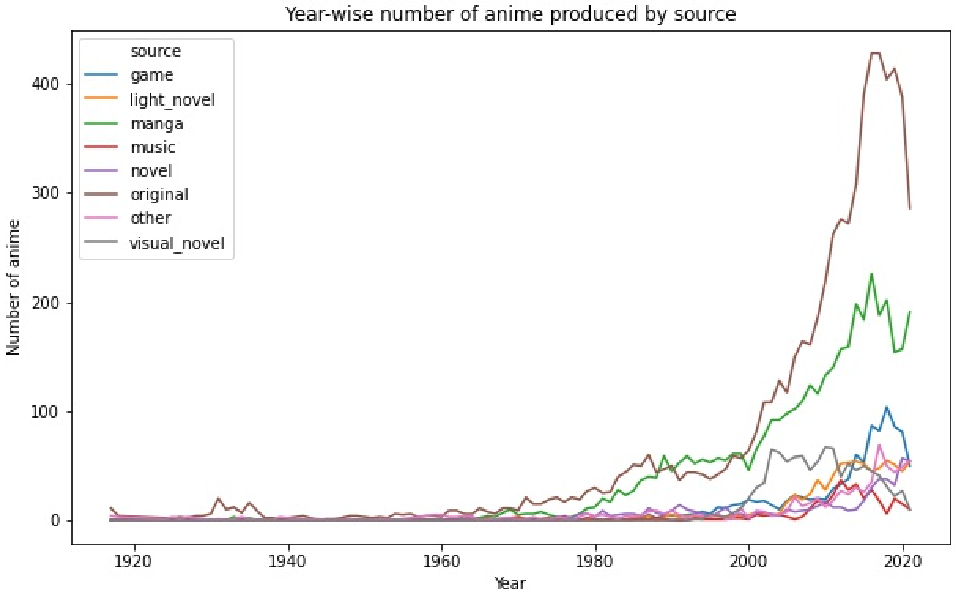

Figure 21: Plot depicting growth of anime production based on their source materials for common sources.

Anime production sourced from original scripts saw very high growth rate in the recent years followed by anime sourced from manga. Other source materials did not see as much growth in their adaptations to anime.

4.8 Length of Anime Trends

We tried to analyze how the popularity and ratings of an anime vary with respect to the number of episodes it has. To do so, we divided the dataset into 9 parts, each corresponding to anime having number of episodes in range $\{0-9, 10-19, 20-29, 30-49, 50-99, 100-199, 200-499, 500-999, 1000+\}$. For each of these episode ranges, we found the total number of anime falling in this range, their average rating and average number of users who have watched them (indicative of popularity). We noticed that as the number of episodes increases, the number of anime falling in that range decreases significantly from 3189 anime having 0 to 9 episodes (Excluding OVA, specials and music) to just 8 anime having 1000+ episodes. The average rating of anime is fairly similar in all categories. In case of average number of users, we note that it is very high for long running anime (having 200+ episodes) which is quite surprising.It is also fairly high for moderate length anime (10-30 episodes) since those are the most common category.

We also plotted the average rating of anime for each of these categories over time. In the plots we notice a fairly steady upward trend in case of anime having short duration (¡ 30 episodes) indicating an improvement in production quality over time.

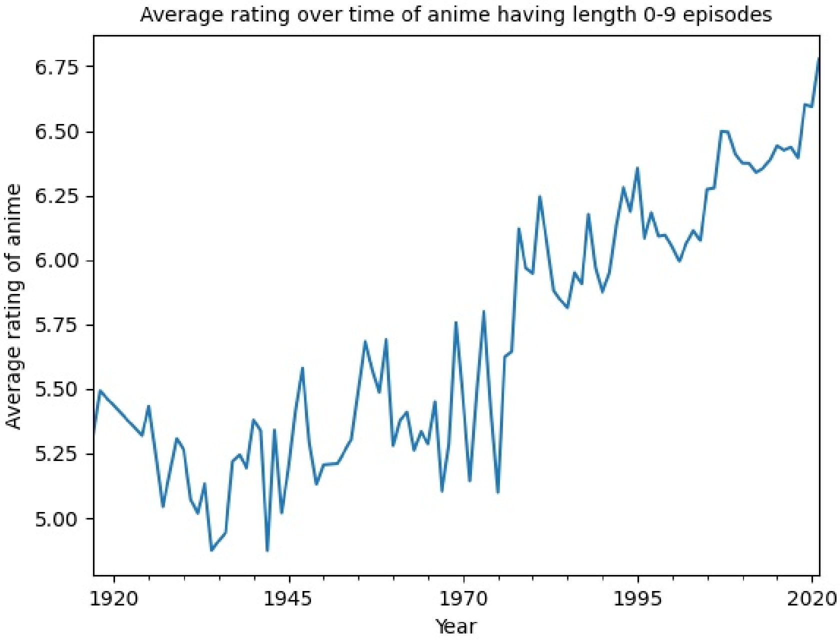

Figure 22: Average rating of anime having length between 0 to 9 episodes

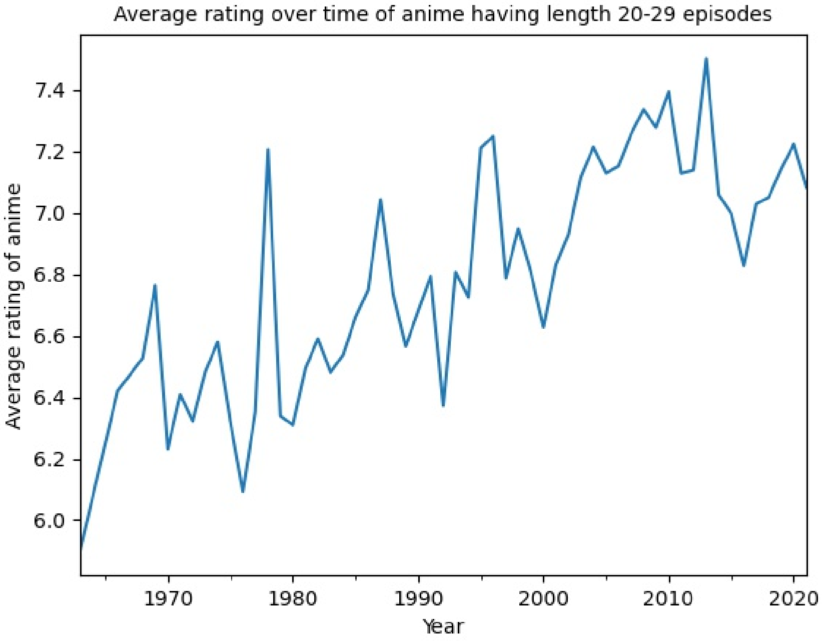

Figure 23: Average rating of anime having length between 20 to 29 episodes

4.9 Distribution of Ratings

We analysed how frequently users rate an anime a particular rating ranging between 1 to 10. On plotting the distribution, we observe a steady upward trend till a rating of 8 where it peaks and then a sudden fall for the ratings of 9 and 10 indicating that the rating of 8 is the most commonly used followed by a rating of 7. This also aligns with what we commonly observe for fairly popular anime since most of them have an average rating between 7 to 8.

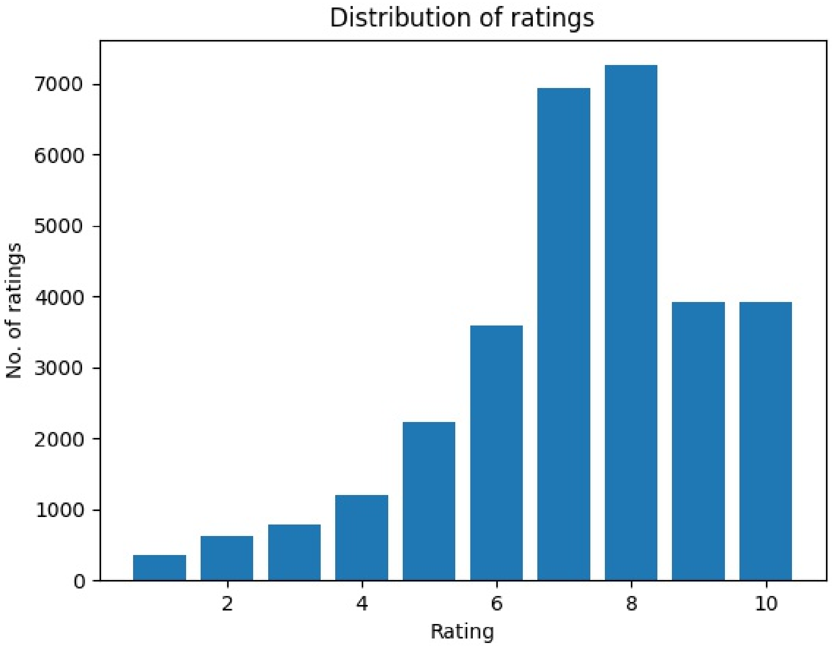

Figure 24: Frequency of various ratings given by users

4.10 Genre Trends

Based on the content or style, an anime can have many genres. We try to analyse what genre is the most prevalent amongst all anime. We formulate this on the basis of most common genres and most watched genres. For figuring out the most common genre, we visualise the count of anime for each particular genre and for most watched genres, we take the average of number of viewers of anime from each particular genre.

4.10.1 Most Common Genre

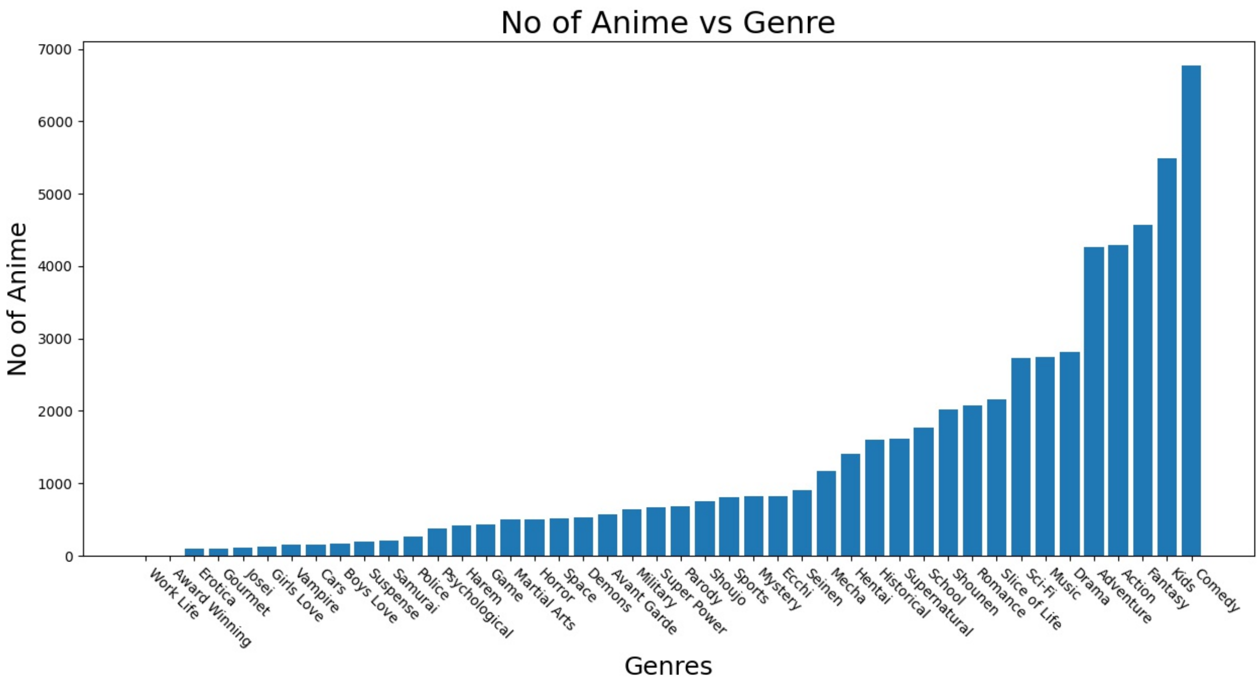

Figure 25: Distribution of number of anime based on their genres

Insights:

1. We find that Comedy is the most common genre. We had a data of 23100 animes and every one in three animes is of the genre Comedy.

2. The top five common genres in decreasing order of number of anime are:

$$\text{Comedy} > \text{Kids} > \text{Fantasy} > \text{Action} > \text{Adventure}$$

3. The five least common genres in decreasing order of number of anime are:

$$\text{Josei} > \text{Gourmet} > \text{Erotica} > \text{Award Winning} > \text{Work Life}$$

4.10.2 Most Watched Genre

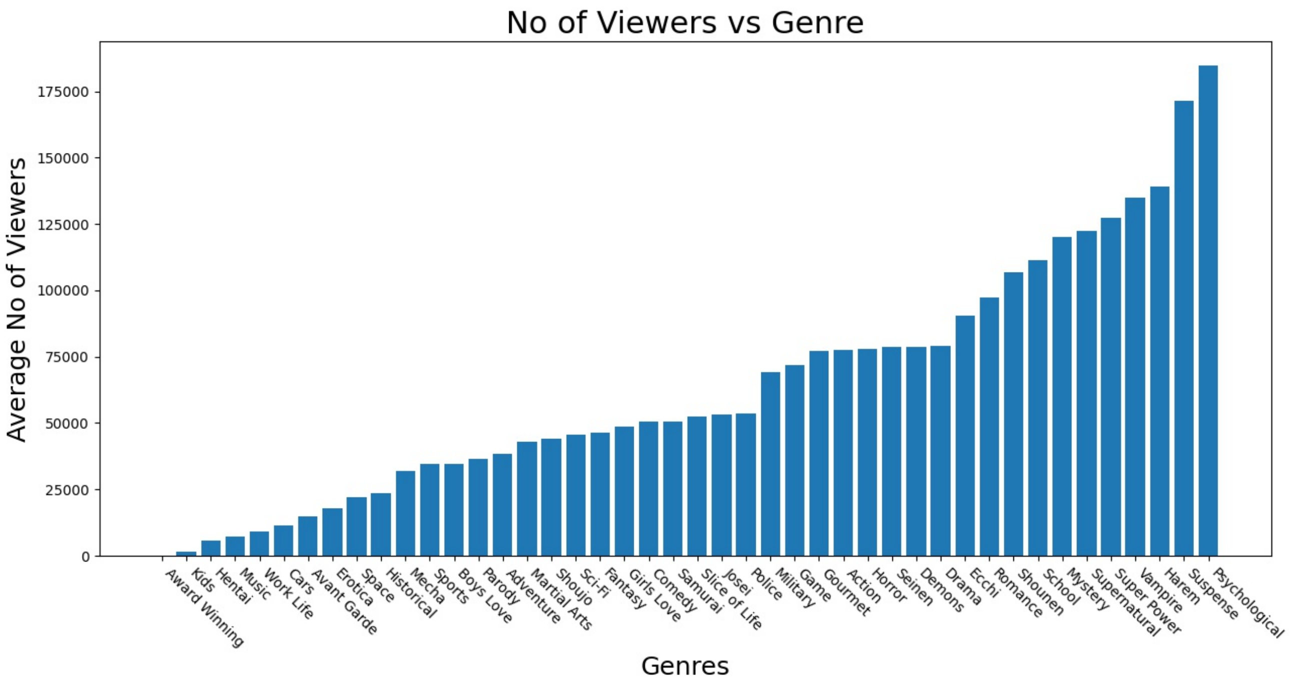

Figure 26: Distribution of number of viewers of anime based on their genres

Insights:

1. The graph of of number of viewers of each genre clearly shows the wide appeal of mystery and suspense amongst the viewers.

2. Genres dominating the anime box office are:

   Psychological > Suspense > Harem > Vampire > Super Power > Supernatural

3. Genres getting least number of viewers are:

   Avant Garde > Cars > Work Life > Music > Hentai > Kids > Award Winning

4.10.3 Analysis of Production Count vs Popularity of Various Genres

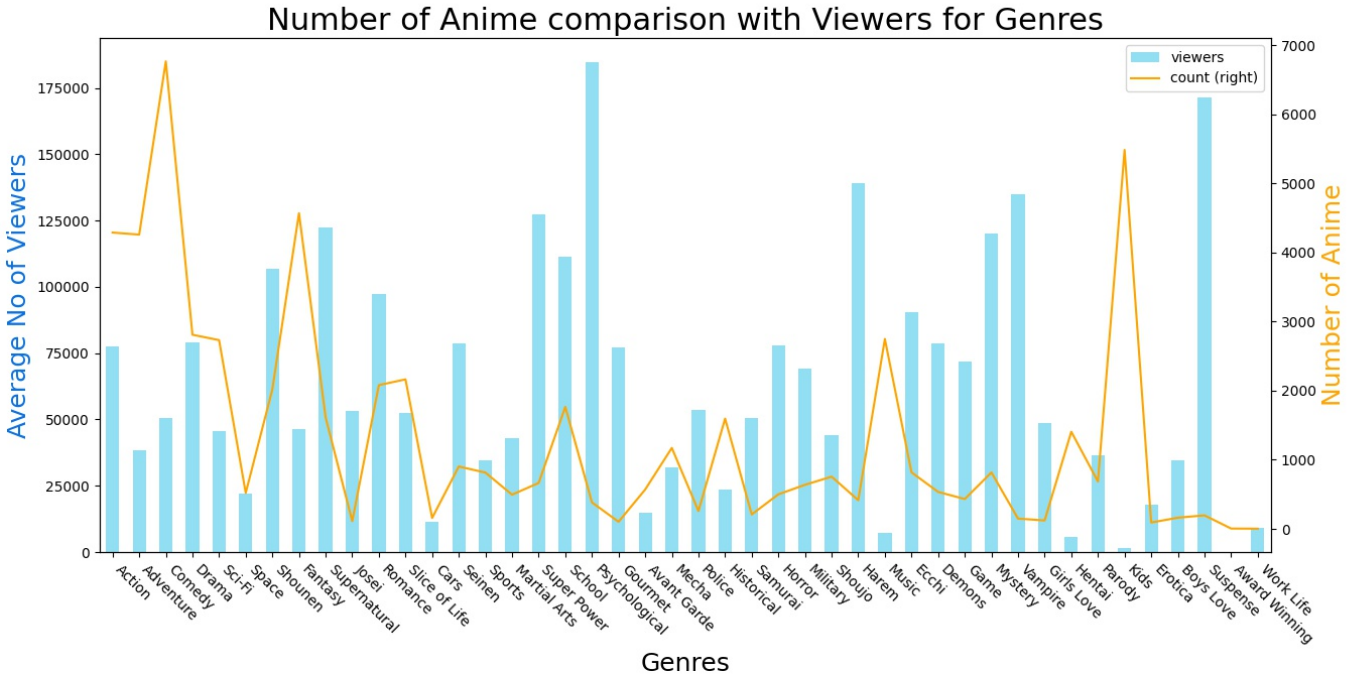

Figure 27: Analysing difference between number of animes produced and total number of viewers of anime of a particular genre

We note that genres like Comedy and Parody which are produced very frequency have significantly lower viewership whereas other genres like Psychological and Suspense which are produced much less frequenty enjoy a very high viewership.

4.10.4 Year-wise Trends in Popularity of Various Genres

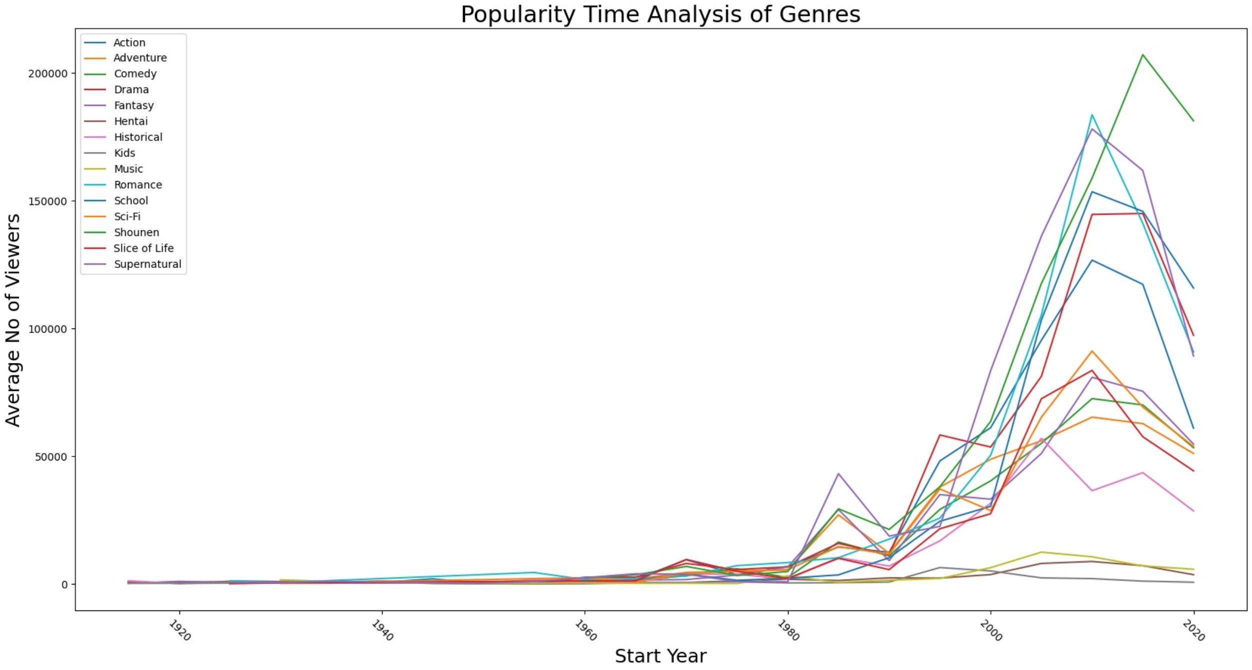

Figure 28: Temporal analysis of number of viewers of a particular genre each year

Here we note a steady rise in popularity of all genres along with some interesting peaks in a few years. We notice that there is a sudden surge in popularity of anime in the mid 1980s especially those from the Supernatural genre. We notice another popularity rise in the mid 1990s after which it has been rising. We can see a sharp peak in the mid 2010s for anime belonging to the Shounen genre. This can likely be attributed to some big shounen titles in recent years such as My Hero Academia and Jujutsu Kaisen.

We also note, like many other plots, the popularity has seen a fall in the recent years. In this case we believe it is not necessarily due to production quality but rather due to the fact that since they have been released recently, not many people have watched them yet and if we look at the plots again a few years down the line then we would expect the average number of viewers for anime in recent years to be much higher.

4.10.5 Year-wise Trends in Production of Genres

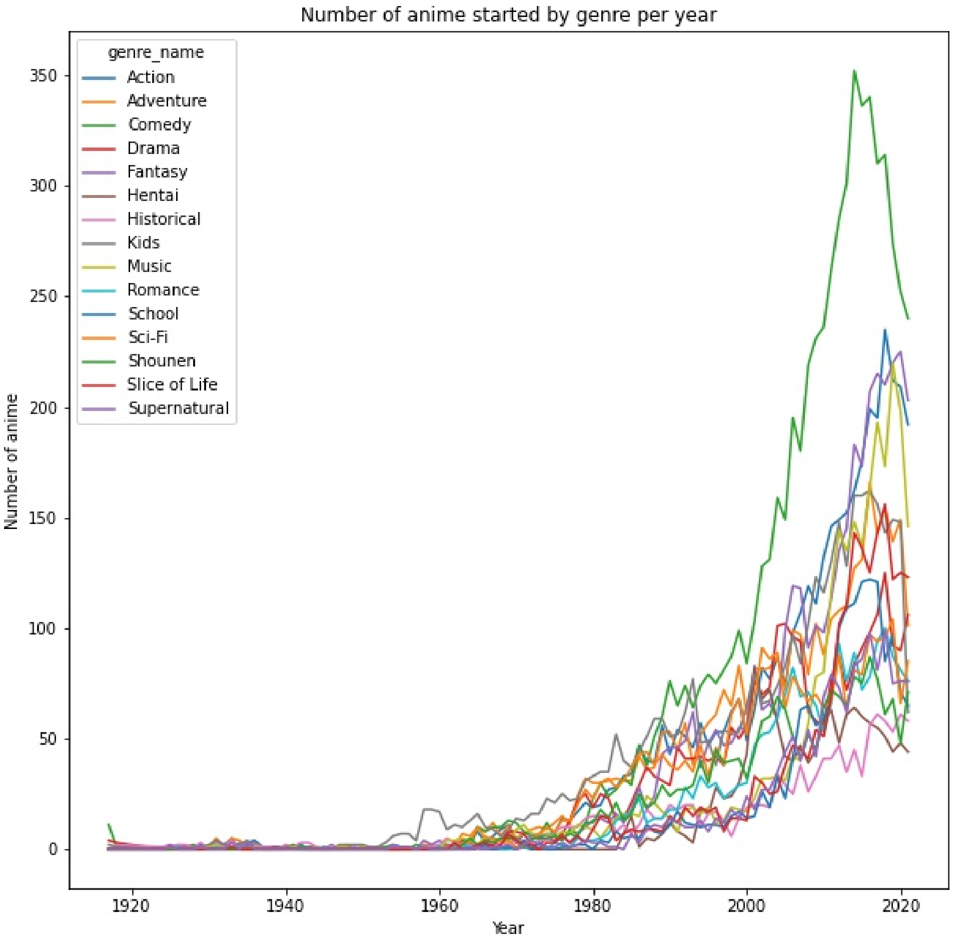

Figure 29: Number of anime started by genre per year

We analyzed the growth of anime by genre over the years. Maximum growth can be seen in the Comedy genre in recent years whereas other genres like Action, Music and Fantasy also saw significant growth. In the plot, it can be observed that the number of anime produced fell considerably after 2020. This can be attributed to the effect of COVID-19, which paused production in unprecedented times.

4.10.6 Mean Popularity and Rank of each Genre

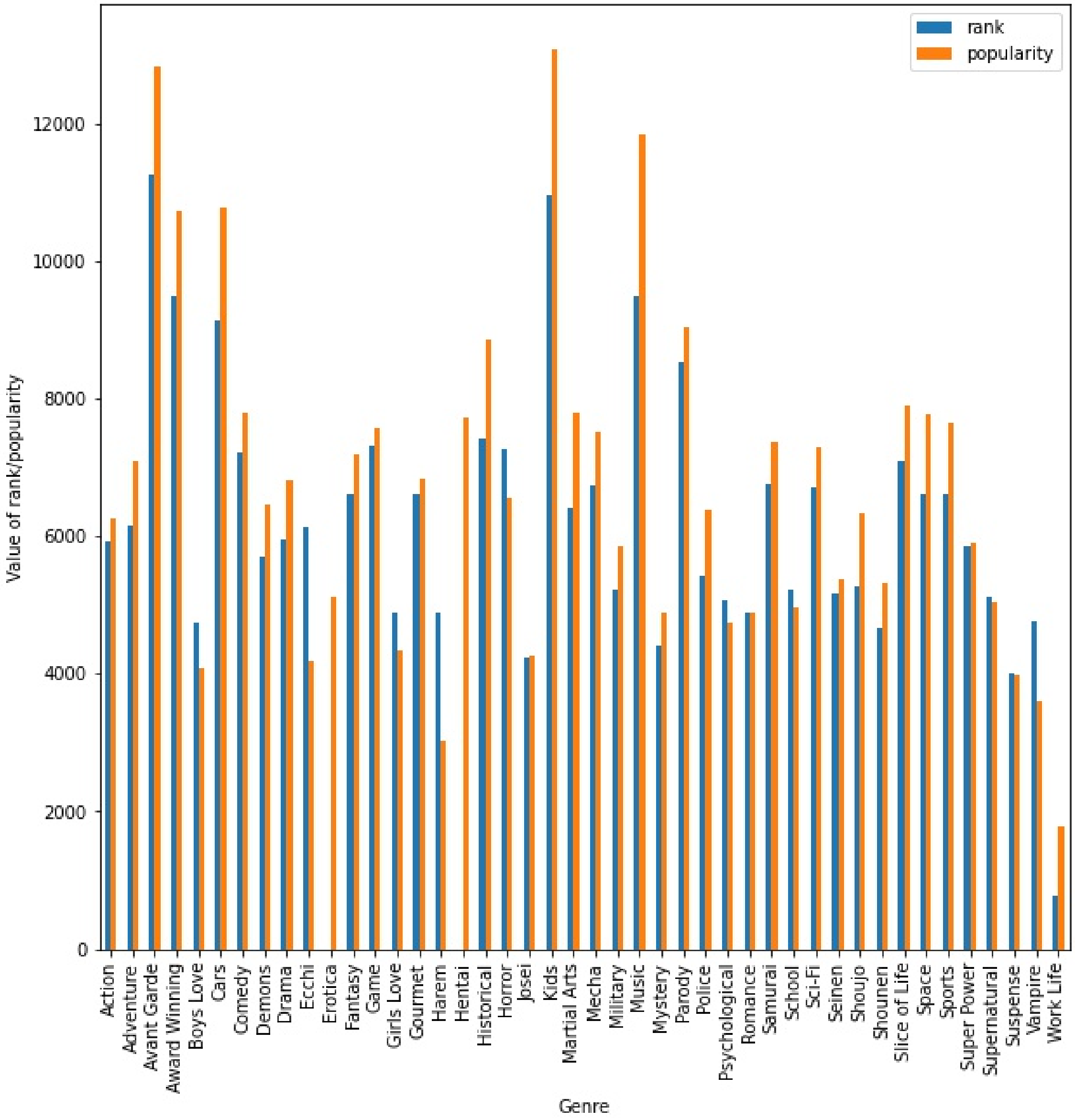

Figure 30: Plot depicting mean rank and popularity for each genre. It is observed that rank and popularity follow similar trend for each genre.

References

[1] Movie recommender based on plot summary using tf-idf vectorization and cosine similarity, Sep 2020.

[2] Dr. Vaibhav Kumar. Singular value decomposition (svd) and its application in recommender system, Jan 2021.

[3] Kevin Liao. Prototyping a recommender system step by step part 1: Knn item-based collaborative filtering, Dec 2020.

[4] Kevin Liao. Prototyping a recommender system step by step part 2: Alternating least square (als) matrix factorization in collaborative filtering, Dec 2020.

[5] Àlex Escolà Nixon. Building a movie content based recommender using tf-idf, Dec 2020.

[6] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In Rudolf Fleischer and Jinhui Xu, editors, *Algorithmic Aspects in Information and Management*, pages 337–348, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.