Problem Statement

Starting in the year 2000 anime, donghua, and Hanguk Aeni — animation originating from Japan, China and South Korea respectively — have been getting an increase in viewership. Newer shows such as Jujustu Kaisen, My Hero Academia, Kaguya Sama: Love is War, and Spy X Family, all garnered popularity from veteran watchers of anime to brand new fans. From the influx of viewership, there has also been an increase of people who sell anime merchandise.

One such store, AniManga Collective, is looking to limit the amount of series they sell, while also maximizing on the anime that is considered popular. Therefore, they requested to compare and determine if certain features make an anime popular. Such features include the genre, the theme, the studio(s), the producer(s), the licensor(s), the source, the rating, the demographic and the viewing type.

Data Wrangling

The dataset found on Kaggle, anime dataset. This dataset was from 1965 to 2022, animes that aired in 2023- 2024 are not a part of this data set. Originally, the anime dataset had 39 columns and around 24,985 rows. To begin, I took any columns off that did not aid in solving the problem — title variations, official artwork banner, synopsis, if the anime was favorited, etc.

Now having 10 less columns, I focused on dropping any rows that have the type ‘music’ as music videos, while visually stunning, are not the focus of the analysis. Next, I drop any rows that have no score, as the score is the target variable. Dropping such values leaves us with 14,464 rows. I fill in any missing value with ‘Unknown’ or ‘Other’. I save this new revised dataset and carry on to exploratory data analysis.

Exploratory Data Analysis

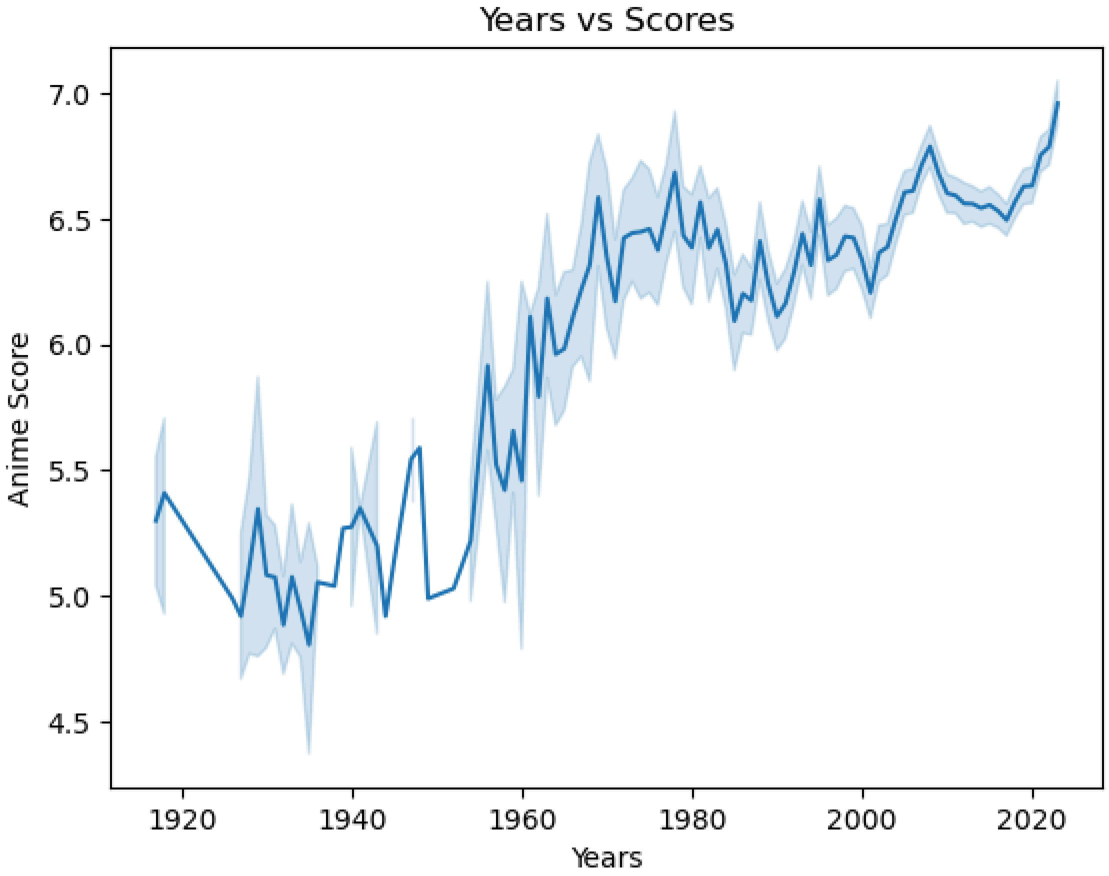
For this section, we start to ask our more prominent questions.

Which year had the most anime premiered?

First, I took the value counts of the start_year column — this column tells us which anime premiered. Then I set up a new dataframe to hold these values. From this new dataframe I take the top ten values, as a result I get the years spanning from 2012 to

2022, excluding 2020, has the years that had the most anime premiered. 2016, being the year the the most anime premiered.

In a similar vein, I wanted to know how the scores might have changed over the years, just because the twenty-tens and the beginning of twenty-twenties were high in the case of premieres, does not mean that the scores are better than older shows and movies. Using a line plot, we can see that the scores actually do increase as the years go by, with 2022 having on average higher scores than years previous.

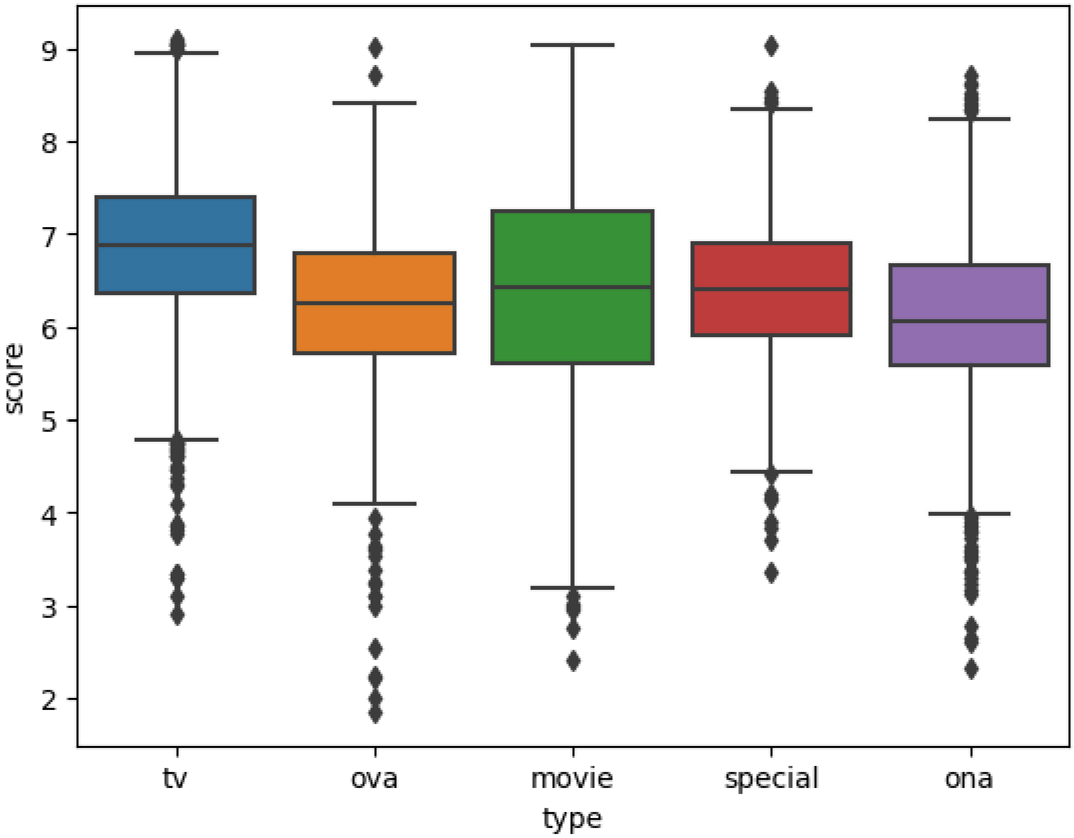


To finish up the time frame of anime premieres, I then looked at the seasons in which anime are released, again just using value_counts, I found that while Autumn is the season with most anime releases, all the seasons are relatively close in amount. With Autumn taking 27.1%, Spring taking 25.1%, Summer taking 24.3% and Winter taking the least with 23.5%.

Which type (Tv show, movie, OVA, ONA) is the most common? Which type has the best scores?

Before getting into the meat of this, I want to go over what OVA and ONA are. OVA stands for Original Video Animation. These types of anime are usually straight to dvd and they were never aired on television nor in theaters if it was a movie. ONA stands for Original Net Animation. These types are animations that are released directly to the internet. ONAs can also be on television, so long as they first premier on the internet. Similar to how we started with start_years, I used the type column and used value counts to see which is the most common way for anime to be viewed. The answer being, TV shows. The type TV show has over 4,000 values and holds 31.5% of the overall type.

To answer which type had the best scores, I grouped the types and used an aggregate function for the score to be at the median. I used the median so we can get a better idea of the distribution, and then we sort the values to be in descending order. Tv shows also had the highest median score of 6.87 and ONA has the lowest of 6.06.

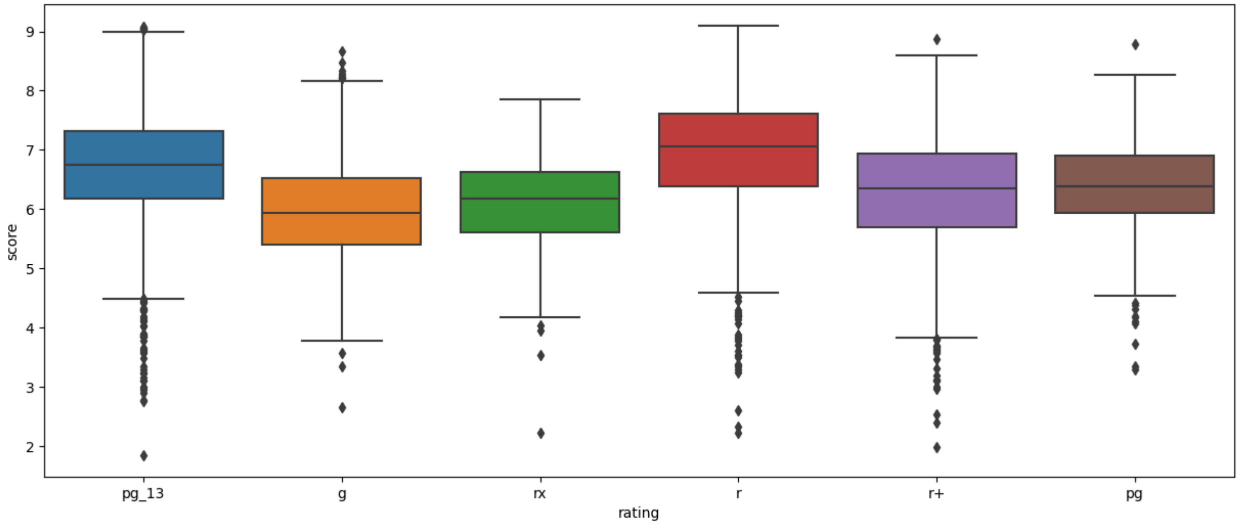


What is the most common anime source? Which source has the best median score?

Finding the value counts of the source, by using value_counts, the data tells us that by a massive margin two sources are held above the other 14, manga and original. Manga having inspired over 4,000, and original having inspired just over 3,500. When looking at the median score for all types we see that web_novels and light_novels have the highest scores with 7.04 and 7.01 respectively. With manga and original having scores of 6.88 and 6.17 respectively. These scores could be skewed due to the overwhelming amount of anime inspired by manga and original, while web_novels and light novels have only inspired 45 anime and 905 anime.

What is the most common anime rating? Which rating has the best median score?

For the most common rating for anime, once again using value_counts, pg-13 takes the cake with over 6,000 anime having a pg-13 rating! The G rating coming is second with just under 3,000 anime, rx — anime that have excessive nudity, sex— has just under 1,500, r — anime that has violence and swearing— has around 1,200 anime, r+— anime that has mild nudity, violence and swearing— have just over 1000, and pg is just under 1000.
When comparing to the scores, R is actually the favored rating with 7.060.



From the box plots we see that R rating has no upper outliers but does have a few lower outliers. RX rating has the least amount of outliers. The G rating has the lowest scores with a median score not quite clearing 6, so maybe around 5.8 or 5.9.

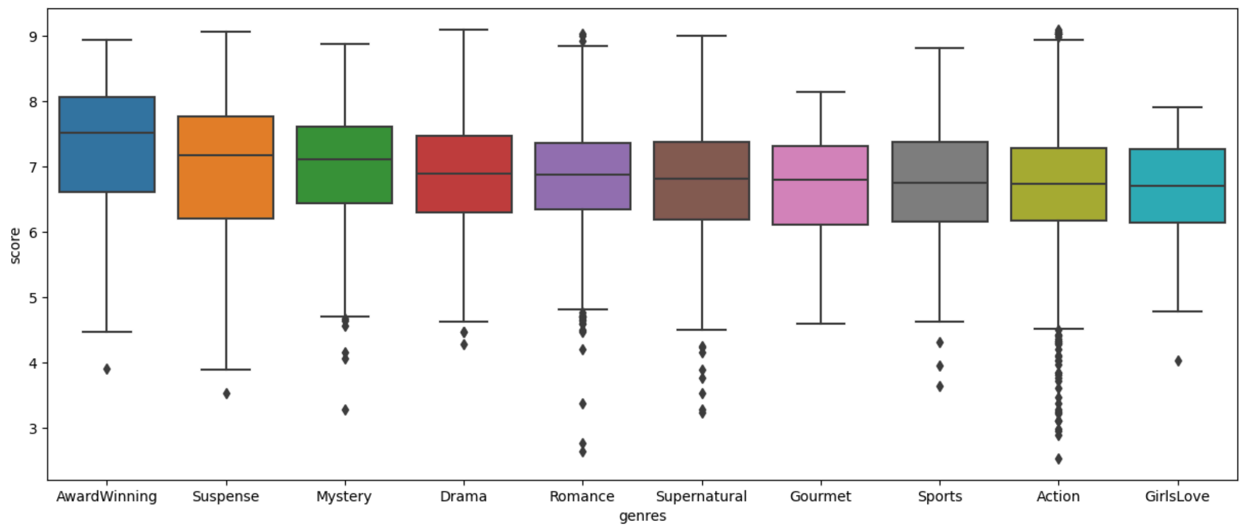
What is the most common anime demographic? Which demographic has the best median score?

The demographics come in technically 5 values, two more values are in the data set as a combination. Those values are Shounen, Kids, Seinen, Shoujo, Josei, (Kids, Shounen) and (Kids, Shoujo) Looking at the demographic counts, Shounen is the most common demographic, then kids, seinen, Shoujo, Josei, (Kids, Shounen), and finally (Kids, Shoujo). If we were looking at percentages, then shounen has 39.8%, Kids with 22.0%, Seinen with 20.2%, Shoujo with 13.6%, Josei with 2.6%, (Kids, Shounen) with 1.2% and (Kids, Shoujo) with 0.5%. Comparing demographics with the scores, shows that shounen is still on top with 7.180. Therefore Shounen is the favored demographic and Kids is the least favored with a score of 6.1.

For the next few features— genres, studios, producers, licensors— is a bit more complex. Since there can be more than one of these features per a single anime— for instance a genre could be a Comedy and an Action series— and I wanted to get an accurate count of how many times a single value showed up. I had to break apart these lists within a series. To do this I set up a function to go through the list of a Series and used the lambda function to split the values wherever a comma appeared.
To get the scores, I had to connect this new DF that had all the unique genres, connect to all the anime that had that genre in it. So a new function was created to take in the Split column and the score column and connect the needed value with the score value. Keep this in mind while I walk you through the rest of the features.

What is the most common anime genre? Which genre has the best median score?

After separating all the genres I took the top ten genres by using value_counts. Comedy was the most common with 5130 anime that had this genre in them. Then, we want to know the genres with the highest score. Using our score connecting function , we find out that AwardWinning, to not much surprise, was the top score, with 7.525. Excluding that score, the next highest score was Suspense with 7.180.



From the box plot we see that Award Winning, Suspense, and Mystery are all above the score of 7, while the rest are all relatively close together, being at the cusp of 6 and 7. If we ignore the Award Winning value, then the favored values are suspense, mystery and Drama.

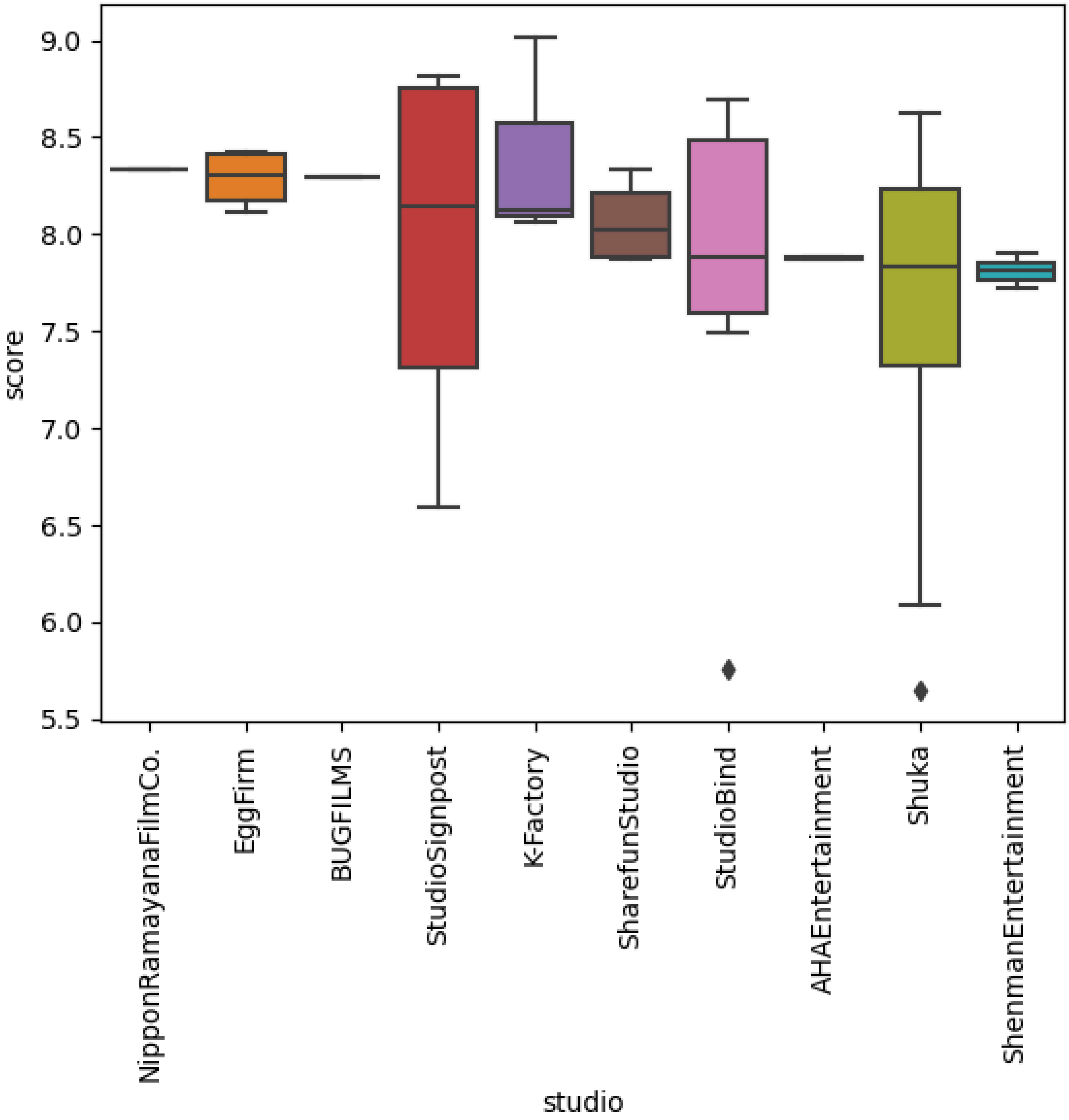
Going backwards a bit, I want to look at the scores for the most common genres. The top most common being, in order are Comedy, Action, Fantasy, Adventure, Sci-fi, Drama, romance, Hentai, Supernatural and Slice of life. Drama, Romance and Supernatural were the highest for the most common genres, 6.90, 6.87 and 6.82 respectively.

What is the most common anime Studio? Which studio has the best score?

Using a similar method for studios as with genres, I took a look at the top ten studios. The studio with the most anime is ToeiAnimation with 723 anime under its belt. As for the scores for the top studios, Toei Animation ranks 9th out of 10 with a score of 6.640. The studio that took the heist score within the top ten is A-1 Pictures with a score of 7.180.

Looking at the studios that have the highest score and would be considered the ‘favored’ studios, almost all of them have a score of 7.8 or higher. The highest being

Nippon Ramayana Film Co with a score of 8.330.



From the box plot, there's three studios that look like there's just a line, which means that these studios only worked on 1 or 2 anime. So I took a look and found that all these studios had less than 20 anime. In fact the majority had worked on less than five. To that measure, I decided that the top ten most common would stand in for what we need, since the top scores studios are skewed. Therefore the favored studio in this case would be A-1 Pictures.

What are the most common anime Producers? Which producer has the best score?

Following up, the top ten most common producers were Aniplex, TVTokyo, Lantis, Movic, BandaiVisual, AT-X, Dentsu, PonyCanyon, Kadokawa, FujiTV. Aniplex had the most anime with 555 and FujiTv had the least with 317.
Dentsu had the highest score, however, with 7.43 and Lantis had the lowest score with 7.01.
For the highest scoring producers, we have MiracleRobo, Annapuru, Sumzap, CDProjecktRed, voqueting, RexEntertainment, AudioHighs, FBC, Funimation, SonyMusicSolutions. With MiracleRobo having the highest score of 8.9. The rest score high as well with all producers scoring 8.4 or above.
I decided to check the value_counts, after all this seems too good to be true, similarly to studios, half of the producers only had one anime under their belt and at most only had ten. So, once again, I would consider that the ten most common is a more accurate measure. Dentsu will be considered the ‘Favored’ producer.

What are the most common anime Licensors? Which licensor has the best score?

Top Ten most common licensors are Funimation(1211), SentaiFilmworks(925), DiscotekMedia(501), ADVFilms(304), MediaBlasters(288), AniplexofAmerica(240), BandaiEntertainment(201), VIZMedia(199), GeneonEntertainmentUSA(181), and CentralParkMedia(154). The scores for these licensors were 7.2 and above. The lowest score being CentralParkMedia, 7.20 and the Highest being Aniplex with a score of 8.16.
Highest scoring licensors were also too good to be true. All the scores were above 8, so I checked the value counts and lo and behold, all except one had less than 100 anime and half had below 10.
Once again, the ten most common will serve as our favored, and Aniplex, which was in both ten most common and highest scored, will serve as our favored licensor.

What are the most common anime themes? Which theme has the best score?

Top ten most common themes are School(1687), Mecha(1048), Historical(936), Music(727), Military(584), Parody(547), SuperPower(534), Mythology(529), Space(453), MartialArts(446), The scores for the top common range from Parody’s score of 6.250 to School’s score of 6.910.

For the top scoring we have AdultCast(408), CGDCT(201), GagHumor(199), Iyashikei(171), Reincarnation(89), OtakuCulture(84), LovePolygon(76), OrganizedCrime(54), Childcare(51) and RomanticSubtext(44). The scores range from OtakuCulture's 7.250 to RomanticSubtext, 7.565.

I'm not thrilled with the fact that more than half of these themes appear less than 100 times. Personally I would prefer using themes that have more than 100 anime to compare scores. If we use the highest score then the favored theme would be 'romantic subtext', otherwise we use the top common highest scoring theme which is school.

Preprocessing and Training

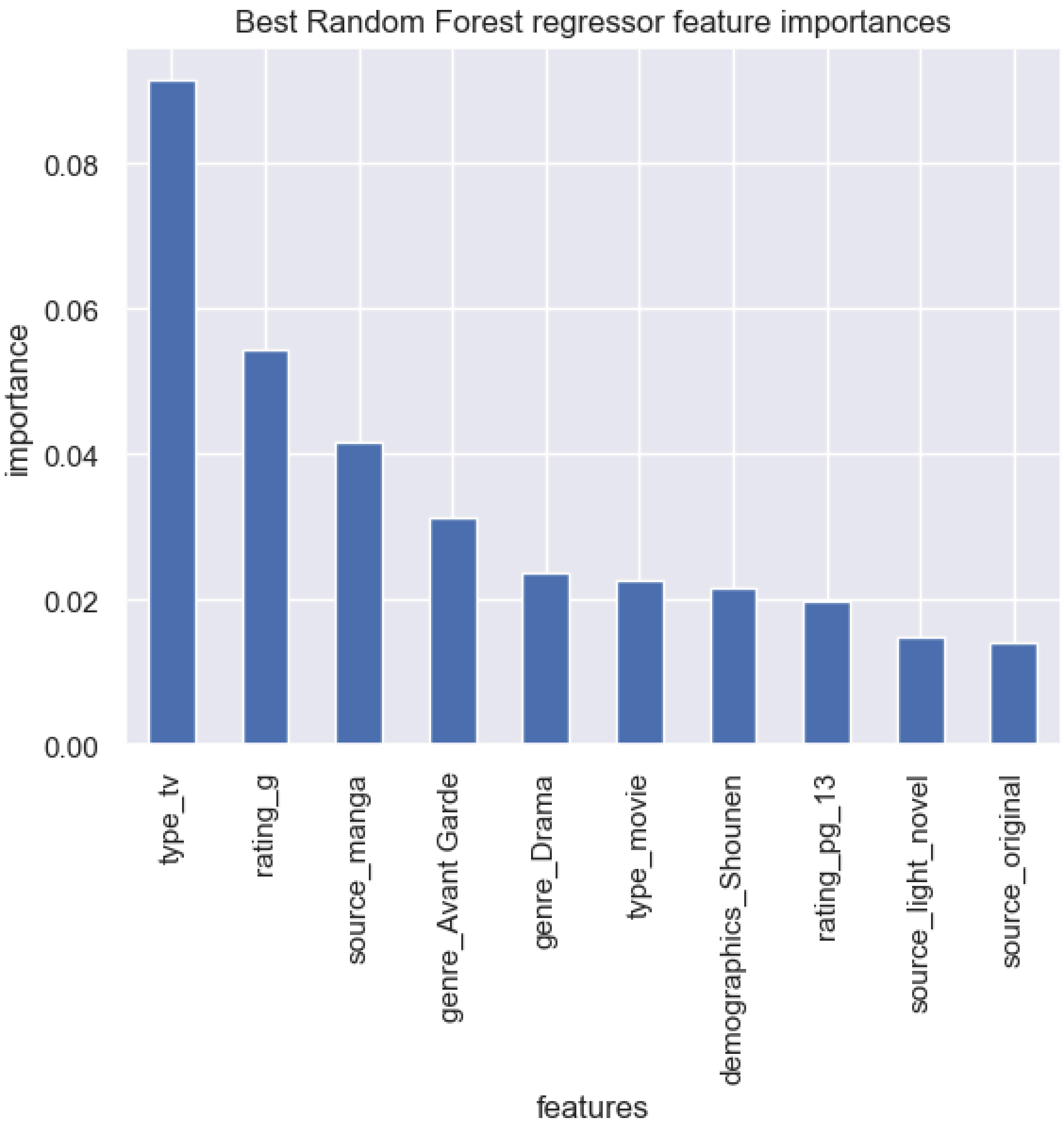
For the training, I needed to get rid of more columns that were not prominent features. I only kept score (our target variable), type, source, rating, demographic, producers, licensors, genres, theme and studios. I split up the listed columns to get an accurate count of genre, producer, theme, licensor and studio. I created dummy variables for all features except the score column.
I separated the features for X. We have all features except score and y is only the score values.

Modeling

For modeling I tested a Linear regression model, a Random Forest Model and a Gradient Boosting model. I base my accuracy on the R2 score.
I made pipes are all models. So starting with the Linear regression model, I used a simpleImputer using the median for any missing values. The base R2 score I received for the training set was 0.687 and the testing test was -1.762. These are not stellar scores, so I wanted to use the optimal amount of features to see if this made a difference. For that we one again make a pipe with Simple Imputer using the strategy mean and SelectkBest with k = 244.
R2 for training set got slightly worse with 0.5236
But got better for test set with 0.4992

Random Forest
For random forest I also used base line pipeline. The R2 scores:
Training: 0.9253
Testing: 0.5590
Looking at our cross validate scores,
Mean cross validation test score: 0.49253300901953717

Mean cross validation train score: 0.5567935438436515
Based on the random forest model, these were the best features.

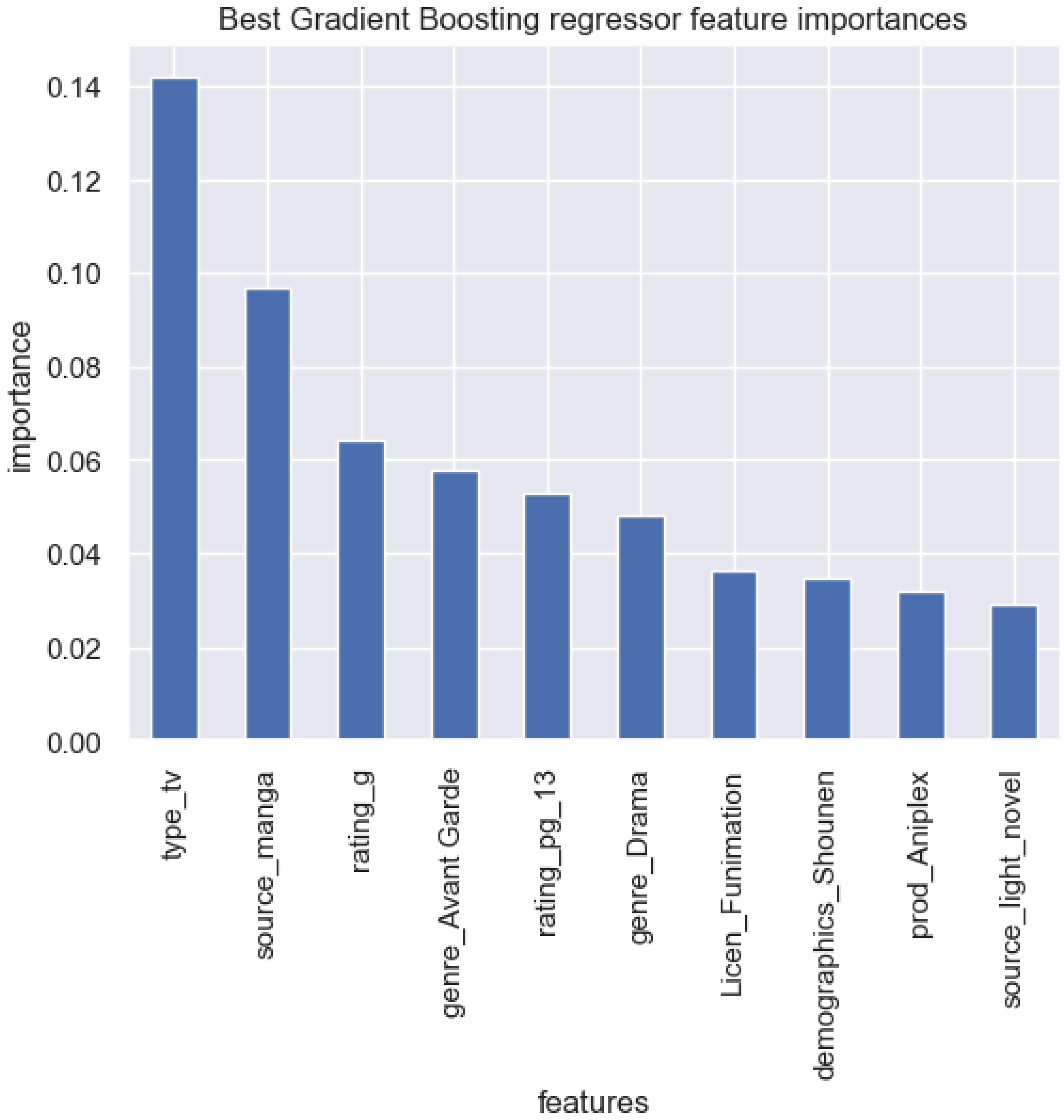


I will not that I did try tuning the Random Forest model, but the scores were actually worse, and I would have tried more combinations of parameter tuning, but the execution time took too long and I did not have the time.
Gradient Boosting

Then we took a look into gradient boosting,

Training set R2 score: 0.48577398169000985
Testing set R2 score: 0.46069815530344305

Mean cross validation test score: 0.4479875912877991
Mean cross validation train score: 0.4490223571670904
Based on the R2 score, Random Forest, even without tuning does better. The cross validation scores are also better in Random Forest.
Gradient boosting important features



Future work

If I had more time, or perhaps I will go back to do more for the modeling stage, I would like to find a way to properly tune the models we certainly have without the runtime going over an hour and a half.

Conclusion

Based on what was found, If aniManga Collective wishes to provide a limited amount of series merchandise, then they should be focusing on animes that are Tv shows, anime that have are rated R, anime that are suspenseful or have some mystery. They should look at anime that have been produced by Denitsu, or look at anime Licensed by Aniplex, Shounen anime, anime with the school theme.