An Analysis of Anime and Anime Consumers Based on MyAnimeList (MAL)

Report by KV Le

CSE 163 Final Project Paper
Professor: Hunter Schafer
Project Mentor: Trinh Nguyen
Project Link: https://github.com/derpyasianpanda/cse163-final

Table of Contents

An Analysis of Anime and Anime Consumers Based on MyAnimeList (MAL) 1

Table of Contents 2

Research Questions: 3

Motivation and background: 4

The Data: 5

Methodology: 5

Phase 1: Retrieve Cleaning Data 5

Phase 2: Analyzing the Data 6

Results 8

Research Question 1: What is an average MyAnimeList user like? What are the differences between a Male and Female MyAnimeList User? 8

Research Question 2: What are the "best" genres for anime? What garners the best ratings and what are the most made genres in the industry today? Has the popularity of genres changed over time? How are they distributed in each time period? 12

Research Question 3: What varies between different animation studios? Is there a "best" studio? 19

Research Question 4: Can we predict the score of an anime on MyAnimeList using factors like genres, animation studio, air date, source material, and etc? What do we change to maximize a score? 31

Challenge Goals: 35

Multiple Datasets: 35

Machine Learning: 35

Messy Data: 35

Work Plan and Evaluation: 35

Testing 36

Collaboration 37

Research Questions:

1. What is an average MyAnimeList user like? What are the differences between a Male and Female MyAnimeList User?

Task Summary:

I will compute the amount averages for MAL users like the number of shows watched, how many planned, how many dropped, ratings, and age. I will also calculate the averages for specific genders and compare Male and Female to see if there are significant differences. I will also combine the anime list, user, and the user lists dataset to find the gender ratios for different genres.

Results Summary:

The Average MyAnimeList user seems to be around 27 years old, being liberal with their scores that average around 8/10, and loves to watch comedy shows. Most users are really into anime, having completed over 100 shows on average and are usually juggling more than 10 shows at a time. There are also pretty significant differences between the genders, with Males generally being “more extreme” with most categories.

2. What are the “best” genres for anime? What garners the best ratings and what are the most made genres in the industry today? Has the popularity of genres changed over time? How are they distributed in each time period?

Task Summary:

Using information about the different animes that have been released, I will be computing many genre statistics, which include the amount of anime made for each genre, most popular genres per year, and the scores of genres for each year. I will also create graphs that display the genre shifts (if any) over the years.

Results Summary:

Comedies are by far the most tagged genre in the industry, with action lagging behind quite a bit. When it comes to “main” genre tags, Action takes a considerable lead ahead of the now second place Comedy. However, the most well rated tagged genre would be Thriller, with Psychological as second place. With that said, Anime as a medium can be seen growing with the amount of genre tags per year increasing. Sadly, there was not enough clear evidence of genre dominance changing over time or having significant distribution changes.

3. What varies between different animation studios? Is there a “best” studio?

Task Summary:

Using information on different animes, I will analyze each animation studio and determine if there are “high quality” studios that people consistently want to see and rate highly. I will attempt to find this through information like studio ratings (for an anime), and genres they may excel in. By using that information, I can compute averages that can help visualize a studio’s “quality”.

Results Summary:

The most highly rated studio is Studio Chizu, which is closely followed by Egg Firm. The worst rated studio was Three-D with Studio! Cucuri being the worst. With the most highly rated genre Thriller, White Fox was the most highly rated Studio and Production Reed was the worst.

4. Can we predict the score of an anime on MyAnimeList using factors like genres, animation studio, air date, source material, and etc? What do we change to maximize a score?

Task Summary:
Using machine learning, I want to take factors like genres, animation studio, and etc, to train a model that predicts how an anime will be rated and received by the MAL community. I will also train another model using similar factors to predict the popularity (through user favorites) of the anime as it releases. This will be done through a Regression Decision Tree Model and with a model trained, I will then try to analyze what makes a highly rated and popular show, determining what factors are most important to an anime's score and popularity.

Results Summary:
The machine learning models were fairly successful with the Mean Absolute Error for score being less than one point off and the favorites being around 800 people off. It seems the most important factors to maximize an anime's score is its duration and studio. For popularity (favorites), it seems the two most important factors are the type of anime (TV, Movie, Special, etc) and the amount of episodes it has.

Motivation and background:

Anime is a medium that has been growing very rapidly, often taking the world by storm with hits like Dragon Ball, Attack on Titan, or Spirited Away. Even large streaming companies like Netflix have started to produce/fund their own anime with shows like Castlevania, B: The Beginning, and Devilman Crybaby. There is no doubt that anime is having a large impact on the global market as shown with the rising popularity of Anime culture, conventions, and shows. However, much of today's anime is made in a traditional sense, meaning animators hand draw every frame of an episode or movie, which can be a long and expensive process. Even with new techniques like flash animation and 3D-Animation, the anime medium has stuck to it's more traditional roots, but that is often what makes it unique. With a lot of room for growth in the anime market, how can studios ensure their future ventures can be successful financially and be well liked?

That is why I decided to pursue this research topic. By analyzing past shows/movies and the demographics behind them, I can find and piece together information, forming patterns that can help predict the success and popularity of an anime. With these predictions it could help animation studios and licensors make better decisions on how to make a commercially successful anime (by measuring popularity) but also a well received anime (by measuring ratings). This can help alleviate the issues of wasted pilot episodes and shows and help animators work on projects that they can be relatively confident that they can be paid for. In addition to predicting success, a model can predict failures so that studios that want to try experimental ideas can get a gauge on how the public might receive certain combinations of genres or studio collaborations. Overall, this information, especially if produced in a high quality manner could help boost efficiency in decisions and create a better anime industry for all.

The Data:

https://www.kaggle.com/azathoth42/myanimelist - *Author: Azathoth*

About the data: This data contains information on the MyAnimeList (MAL) Database up to June 2018. MyAnimeList is a site that contains a wealth of data on Japanese animation and on the users who choose to record the anime-watching progress on it. From anime to user anime lists, MAL contains almost every anime due to its "open-source"-esque nature and a large part of anime consumer information. Azathoth's Kaggle dataset is one that has gathered the information within MAL that "contains information about users (gender, location, birth date etc.), about anime (airing date, genres, producer…) and anime lists." Azathoth has also compiled "cleaned" data that has filtered out much if not all of the nonsensical data to help with any data science endeavors and put them into CSV files.

I will also use an unofficial API (https://jikan.moe/) with a Python wrapper (https://github.com/abhinavk99/jikanpy) to retrieve 25-100 animes after 2018 as those can provide good tests for the machine learning model.

Methodology:

Phase 1: Retrieve Cleaning Data

In the first phase, I must retrieve and clean the data into smaller and more usable formats. To do this, I will be using the pandas library. I will start out by cutting out unnecessary columns in each Kaggle dataset, only retaining the relevant columns for this experiment. The columns I will keep in each are:

Anime Information: anime_id, title, image_url, type, episodes, duration_min, score, scored_by, rank, popularity, members, favorites, related, studio, genre, aired_from_year, source

User List: username, user_id, user_watching, user_completed, user_onhold, user_dropped, user_plantowatch, user_days_spent_watching, gender, location, birth_date, stats_mean_score, stats_episodes

User Anime Lists: username, anime_id, my_score, my_status, my_watched_episodes

Then I will proceed to retrieve anime information after 2018 by using the unofficial MAL REST API called Jikan. There is a python wrapper that can be used to retrieve relevant information that corresponds with the information we kept for Anime Information. Using this I can gather the information I want from a relatively messy data source and organize it in a way that can fit into the machine learning model later on.

After retrieving all the information and cleaning it of all the unnecessary columns, I will then save them into CSV files that can be used in the later phases.

Phase 2: Analyzing the Data
All visualizations will be done with Seaborn, GraphViz, and MatPlotLib libraries. All calculations will be down with the Pandas and internal libraries.

Research Question 1 - For the first research question, I will be analyzing users of MyAnimeList. After using pandas to calculate things like the average age, shows watched, shows dropped, shows planned, and ratings I will then plot them out individually using the sensible encoding formats. I will also calculate those averages for different genders and plot them with bar graphs to see if there are significant differences. After that, I plot the average time spent on Anime depending on age and plotted that out in a scatter plot that is also separated by gender. Lastly, I will join the anime information, user's anime lists, and user lists to find out the gender ratio/distribution for the top genres by finding if a certain user is watching a show.

Research Question 2 - For the second research question, I will be analyzing genres and how they change over time by using the anime list info. First, I will calculate the amount of each genre made in each year by counting up the animes that have a certain genre and then plot many genres and how they've changed over time. Second, I will calculate the most popular genres by their counts. Third, I will plot the most popular "main" tagged genres. Fifth, I will calculate the average score for each genre to see the most well received genre and plot out how it compares with every other genre.

Research Question 3 - For the third research question, I will be analyzing anime studios using the anime list data. After using pandas to calculate the average rating a studio receives I can plot the top anime studios overall. I will then compute the studio's genre average (the average score they recieve for a specific genre) and see which studios top which genres by visualizing the averages for a few genres. Lastly, I will compute the average scores on a yearly basis and plot the top anime studio in terms of reception and popularity for each year to see if there is studio dominance.

Research Question 4 - For the last research question, I will be creating a machine learning model to predict an anime's score and popularity. To do this I will be using the SciKit-Learn library to train a Regression Model to predict the scores from data found in the anime list. I decided to go with a Decision Tree and the labels will be the anime's score and popularity (favorites) and the features will be the following: type, episodes, duration_min, studio, genre, and source. The reason I don’t use all the information in the cleaned or original anime data is because they are features that have no ties to score in this context. For example, the anime’s name wouldn’t be that relevant to the model and it would most likely be detrimental because anime names can be wildly varying and have no tie to quality. Also things like “currently airing” are irrelevant in this context because this data is not completely up to date, meaning many of the animes don’t have the most accurate information in those categories. After training a proper model, I will use the data retrieved from the Jikan API as a test set for the model to calculate its accuracy. After that, I will see what parameters impact an anime’s popularity and score and try to determine what factors are most impactful to an anime’s score and popularity.

Results

NOTE: For all these visualizations enlarged, you can view/download them all here https://drive.google.com/drive/folders/1tGukUVRHnCoPs_m2K7dli0XYCPlIjDPY?usp=sharing

Research Question 1: What is an average MyAnimeList user like? What are the differences between a Male and Female MyAnimeList User?

"The Average MyAnimeList user seems to be around 27 years old, being liberal with their scores that average around 8/10, and loves to watch comedy shows. Most users are really into anime, having completed over 100 shows on average and are usually juggling more than 10 shows at a time. There are also pretty significant differences between the genders, with Males generally being "more extreme" with most categories."

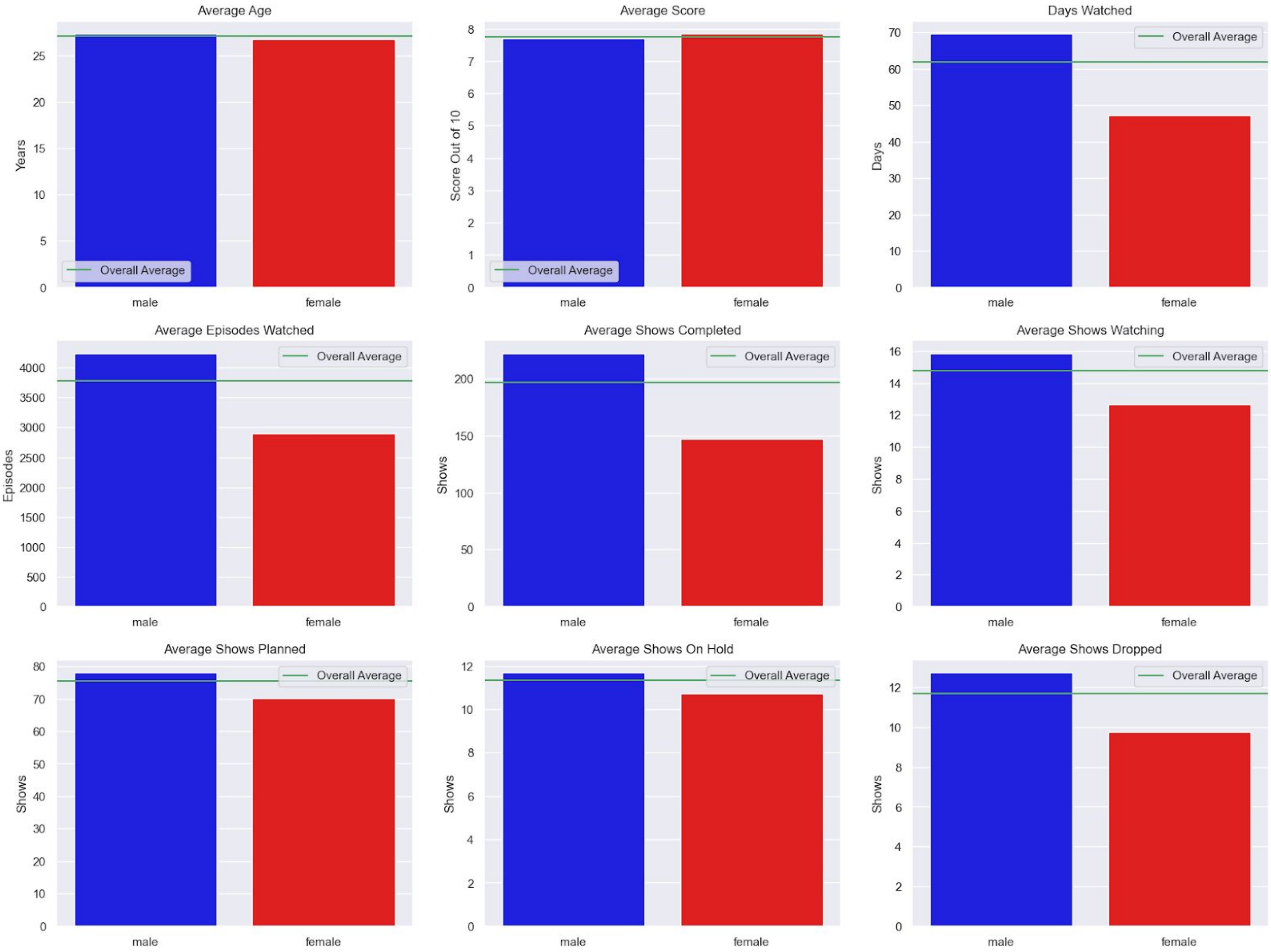

Figure 1.1: Average statistics for MAL users

With the plots that show user averages (Figure 1.1), we can see anime watchers, MAL users specifically, have a real passion for the art. On average, watching more than 3500 episodes, finishing 200 animes, currently juggling around 15, and planning to watch 70 shows later clearly proves the dedication of anime fans in general and the staying power of anime as a medium. It reinforces the notion that anime is and can continue being a global market force/entertainment medium, helping any future investors/producers rest easy that there is still an audience.

When we dive deeper into how the genders are different we start seeing how males seem “more extreme” with the medium. Males watch around 4000 episodes on average compared to the ~3000 episodes by females and this is a similar trend seen with all the statistical averages but scores. Even then, with scores, it can be argued that rating shows more critically shows a “more thoughtful rating system,” though it can also mean females have “better taste” in what shows to watch.

Back to general statistics, I would like to talk more about the average ratings for users and the seemingly large amounts of anime an average user watches at a given time. For average ratings, I found it to be extraordinarily high with averages reaching 8/10. I would have expected it to be near 6/10, but I predict that the average rating might be so high because every user of MAL has a site to help them find better anime, therefore avoiding lower rated shows. I also thought a very interesting average was the amount of anime a user will watch at a single time. Juggling 15 shows at a time is something I find very shocking because that means a user keeps track of 15 different plotlines and tries to keep up watching all of them. I think this is due to “in season/currently airing” anime being so popular because of all the “community buzz” when a show releases. However, I still find it shocking people on average watch 15 shows at a time.

With this project, I also visualized the amount of time spent watching anime, separated by age and gender. I still find it surprising the amount of anime watched is measured in days because it means just average anime fans have already watched well over 24 hours of anime. As seen in the plot below (Figure 1.2), we can see a point for each MAL user that is placed according to their age and time watched, and being colored by gender. When looking at both genders, it seems the age that watches the most anime seems to be around 28-32 years old. However, when we separate by gender, it seems Females watch more anime at a younger age at around 25-30 years old compared to the 28-31 of Males. Some last observations I made was that the Male curve is higher than the Female one (as expected from the averaged before), but also that the Female curve seems more evenly distributed for each age. The second observation might indicate Female audiences are more diverse in their age on average.

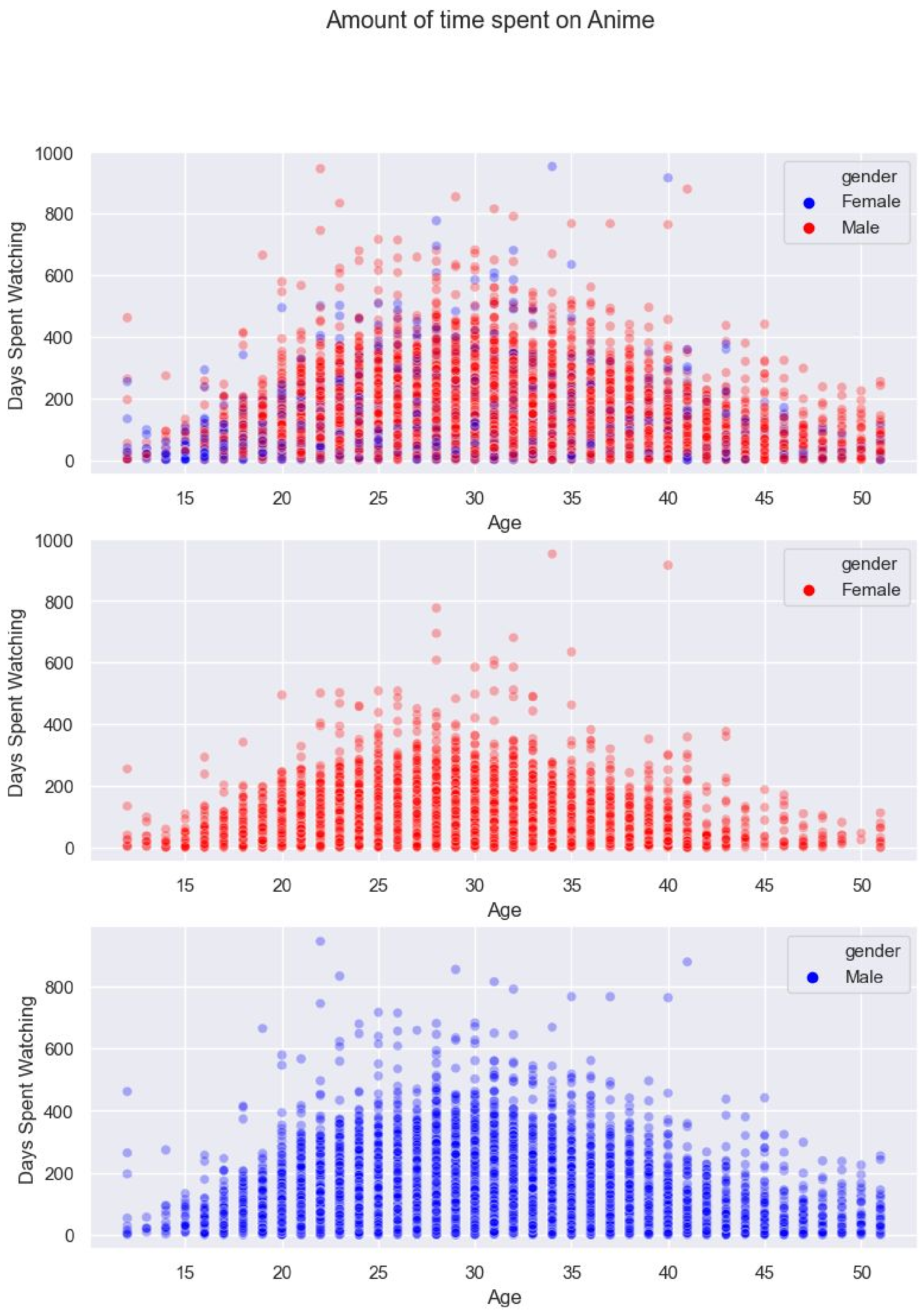

Figure 1.2: Scatter Plots of Average Time Spent Watching Anime

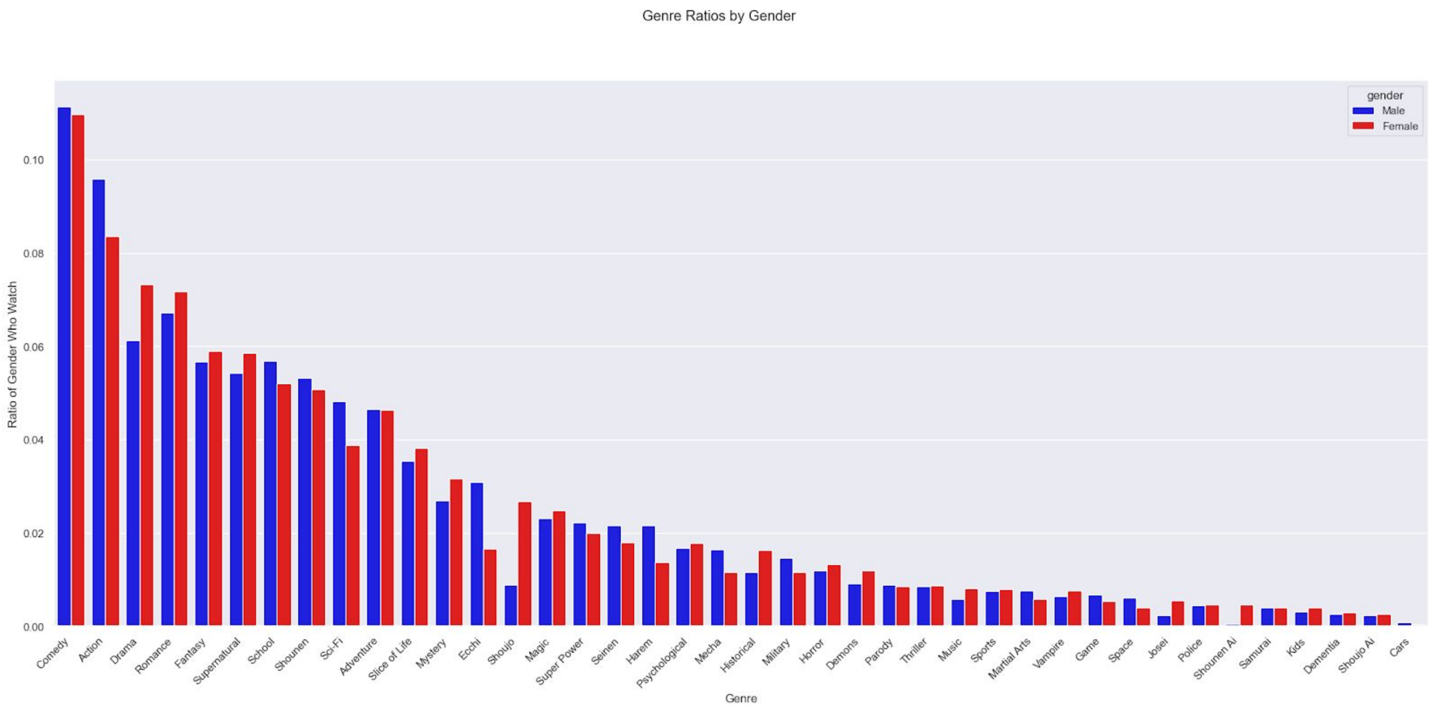

Figure 1.3: Bar Plots of Gender Ratios of Genres

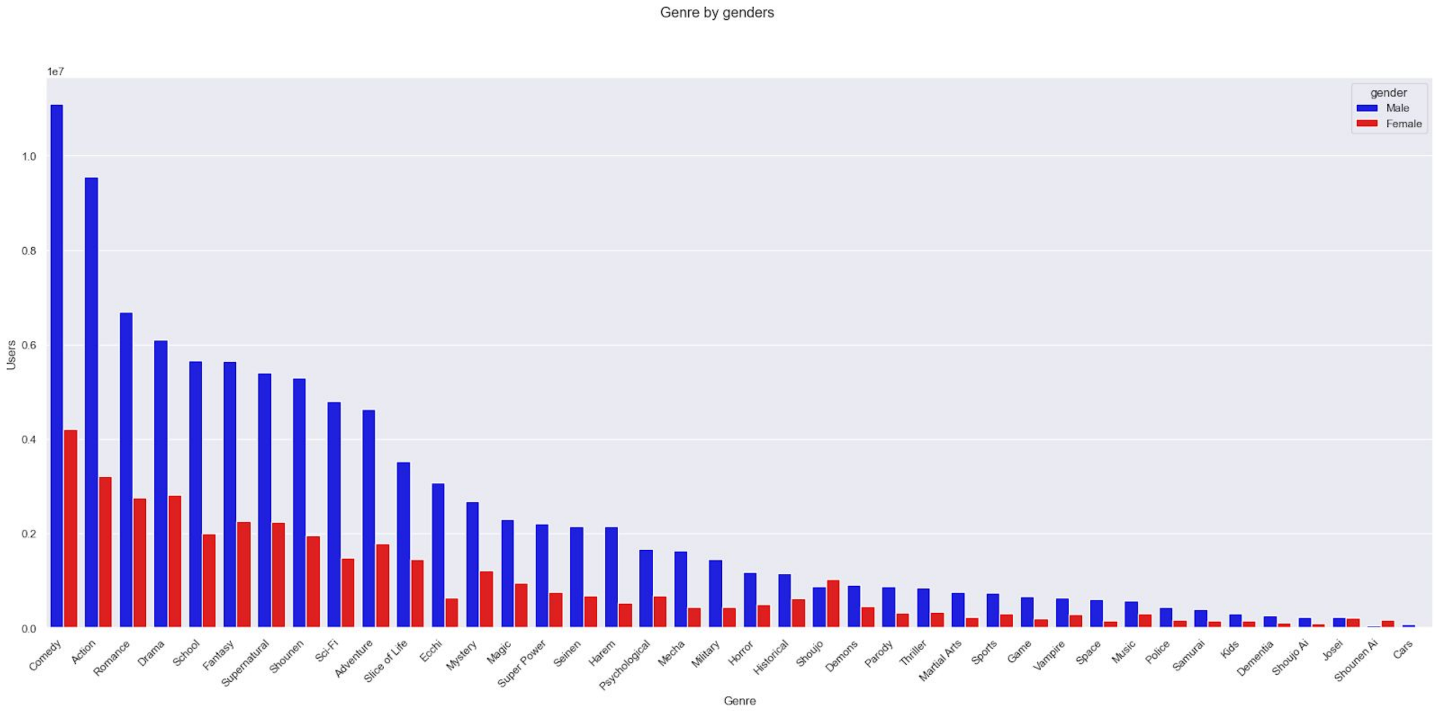

Figure 1.4: Bar Plots of Genre Popularity Separated by Gender

The final visualizations I did were about genre popularity separated by gender. What I discovered is that Males have an overwhelming majority in each genre except Bishoujo (Cute

Girl Anime) and Shounen Ai (Gay Love Stories). However, when calculating ratios (amount watched / total gender count), we find a more compelling story. Females are more into genres like Drama, Romance, Fantasy, and Music while males watch genres like Comedy, Action, and Harem. With ratios, it is easier to see what genres different genders like and is more fair to gauge interest.

What this information proves to studios, licensors, and producers can target certain genres at genders they know receive it well on average. They can also take advantage of how users watch a lot of anime, and proceed to optimize their output based on that. Lastly they can take this information and tailor their shows to a gender or a general user base.

Research Question 2: What are the "best" genres for anime? What garners the best ratings and what are the most made genres in the industry today? Has the popularity of genres changed over time? How are they distributed in each time period?

"Comedies are by far the most tagged genre in the industry, with action lagging behind quite a bit. When it comes to "main" genre tags, Action takes a considerable lead ahead of the now second place Comedy. However, the most well rated tagged genre would be Thriller, with Psychological as second place. With that said, Anime as a medium can be seen growing with the amount of genre tags per year increasing. Sadly, there was not enough clear evidence of genre dominance changing over time or having significant distribution changes."

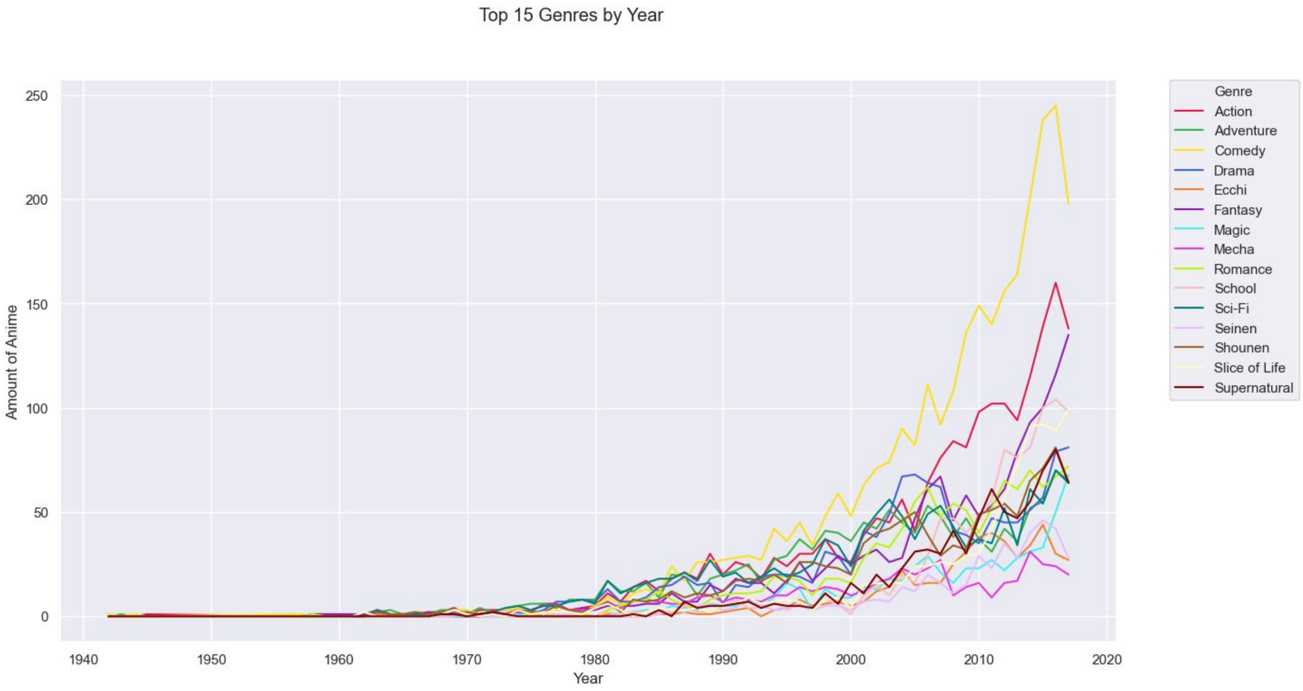

Figure 2.1: Line Plot Amount of Genres Made per Year

In the line plot above (Figure 2.1), it can be seen that Comedy the most popular genre by a pretty significant margin. I was surprised because with a male majority audience, I thought that something like Adventure or Action would be the most popular genre. However, I realized that when tagging multi-genre animes, most animes could be tagged as Comedy but have a focus elsewhere. This doesn't take away from Comedy's popularity though as I believe the amount of anime tagged with it is a testament to the genre's prevalence throughout the industry.

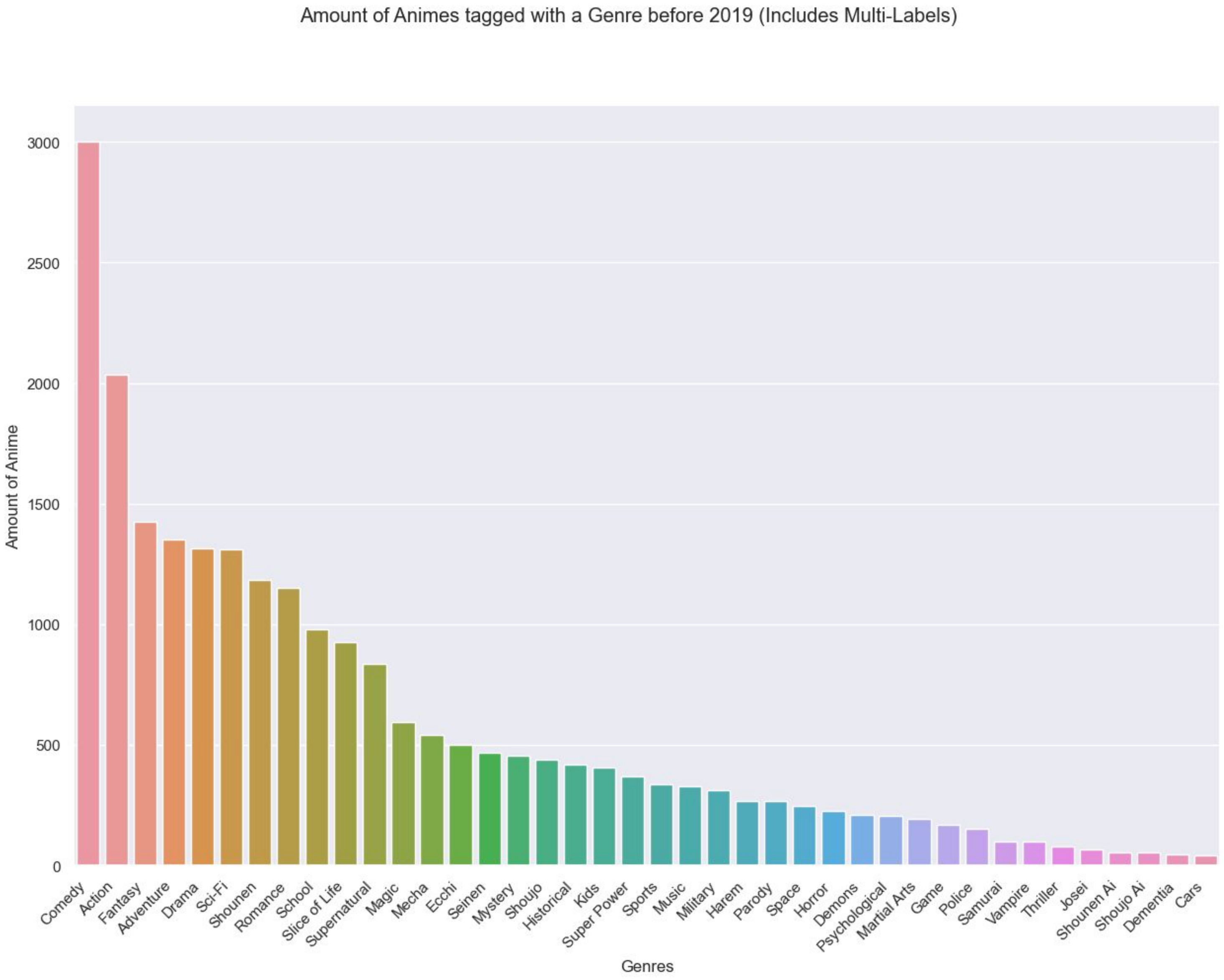

Figure 2.2: Amount of Animes Made Tagged with a Genre (Before 2019)

Figure 2.2 further solidifies Comedy's relevancy within the anime industry and shows that it is a genre that people will see (but not necessarily rate high). With a towering 3000 animes with a

Comedy tag and Action genre lagging behind with 2000, Comedy is the most popular genre within anime. It can also be seen that other genres don't have as much distribution for their tags.

However, Comedy doesn't take a lead when it comes to a main genre tag. According to Figure 2.3, Action takes the trophy for being the most prevalent main genre for anime. Comedy is now in second place, which is still a testament to its popularity. But having Action be the most popular main genre indicates that Action is the most focused aspect of the anime industry. This can be seen with some very culturally relevant animes like Dragon Ball or Death Note.

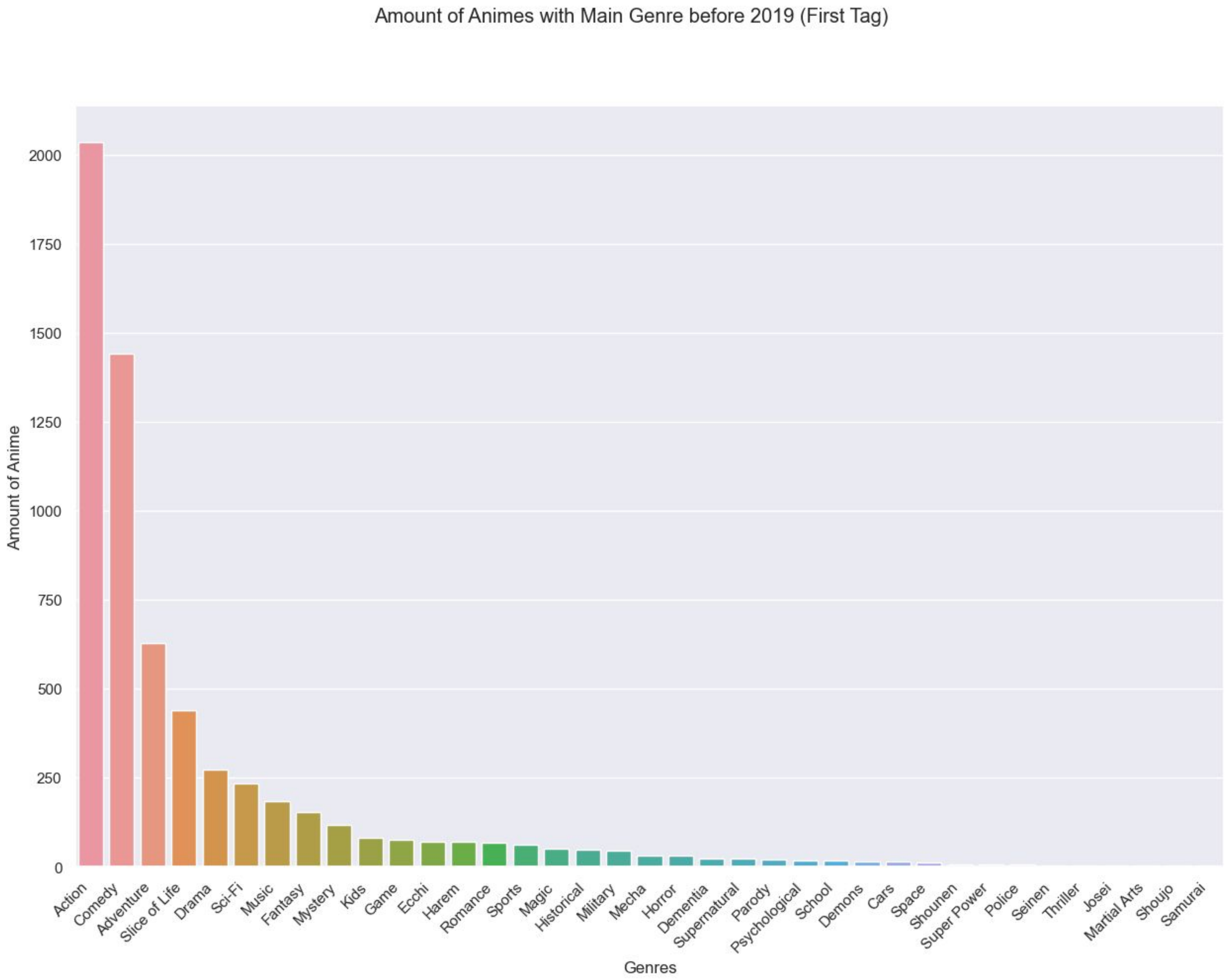

Figure 2.3: Amount of Animes Made Tagged with a Genre as its Main Genre (Before 2019)

Despite Comedy's permeation throughout the industry, it might be dethroned as the most multi-tagged genre in a few years because in the Figures 2.4 and 2.5 below, the latest 2019 data shows that Action tags have taken the lead in both multi-tag and main-tag. While not by a huge margin, Action has overtaken Comedy in 2019 and if this continues for the coming years, Comedy might be dethroned.

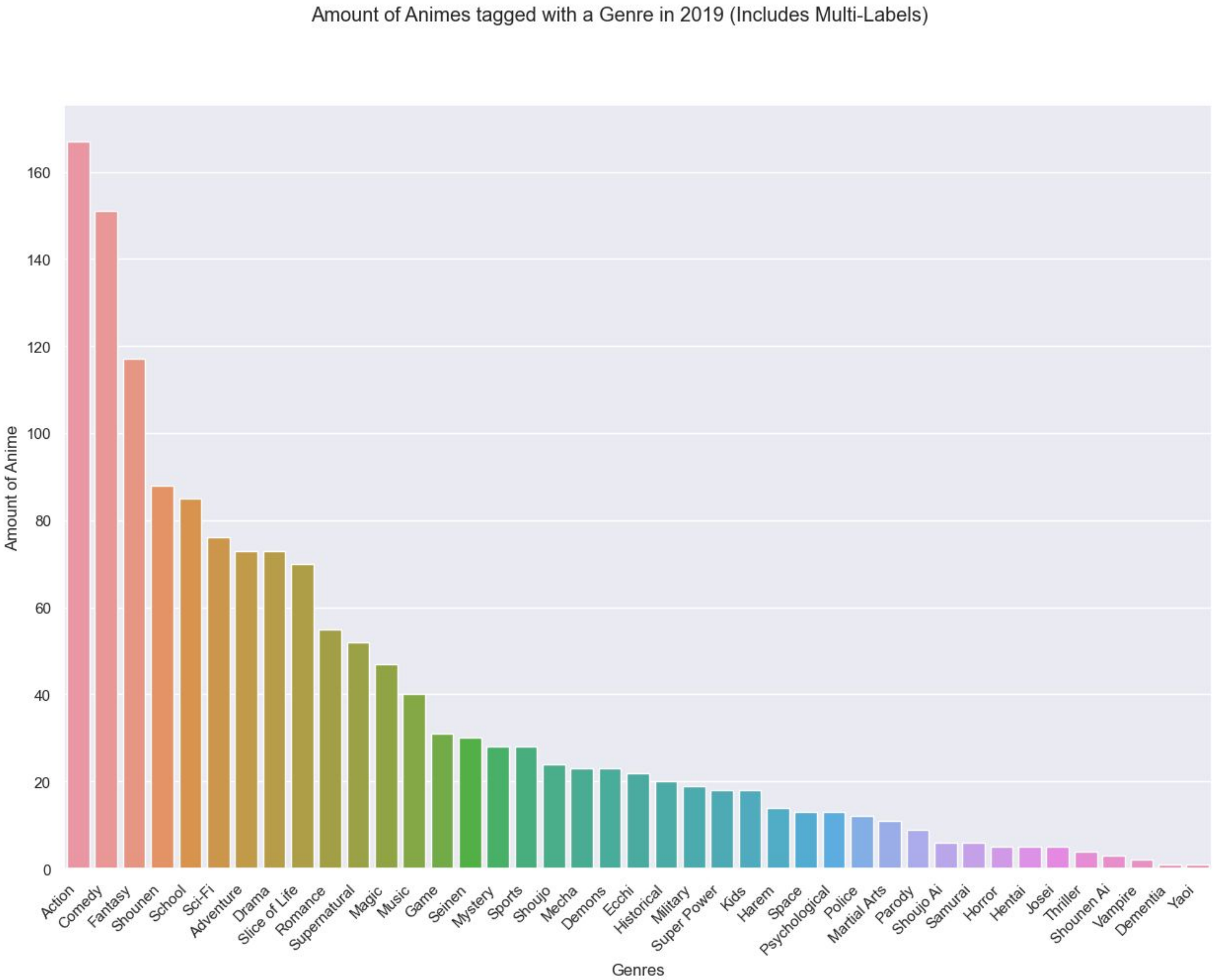

Figure 2.4: Amount of Animes Made Tagged with a Genre (In 2019)

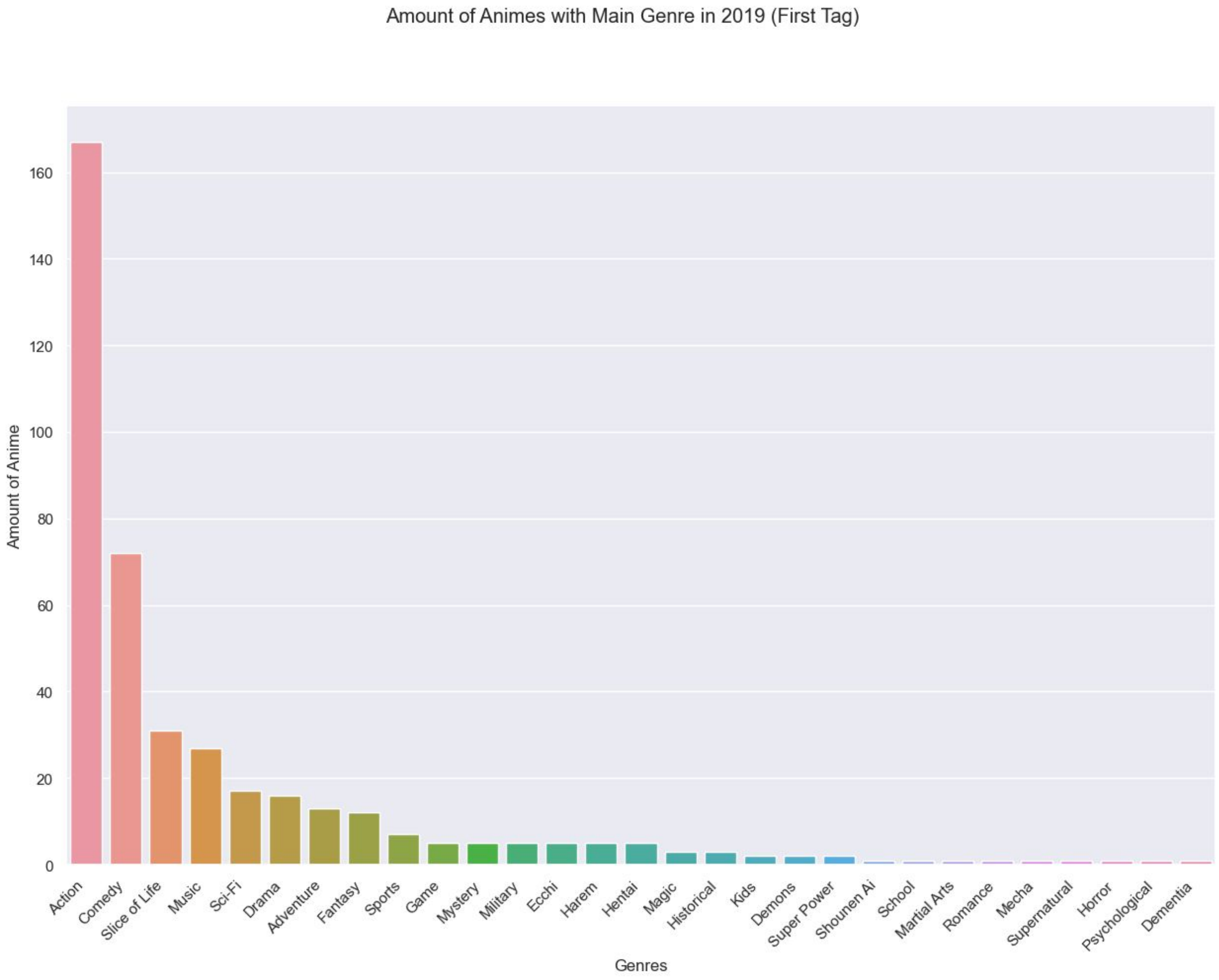

Figure 2.5: Amount of Animes Made Tagged with a Genre as its Main Genre (In 2019)

For the most well received genres, Thriller is the most highly rated with an average of around 7.5/10 as shown below in Figure 2.6. Psychological follows closely, averaging around 7.3/10. Comedy and Action are both at or below the halfway point being relatively low. Even with industry focus those two are most likely brought down by the many failed attempts to capitalize on those genres. This can be further backed with how Thriller currently (2019) has only 122 anime listings with Comedy at over 5000. In a sense, “Quality over Quantity” applies to anime genres too.

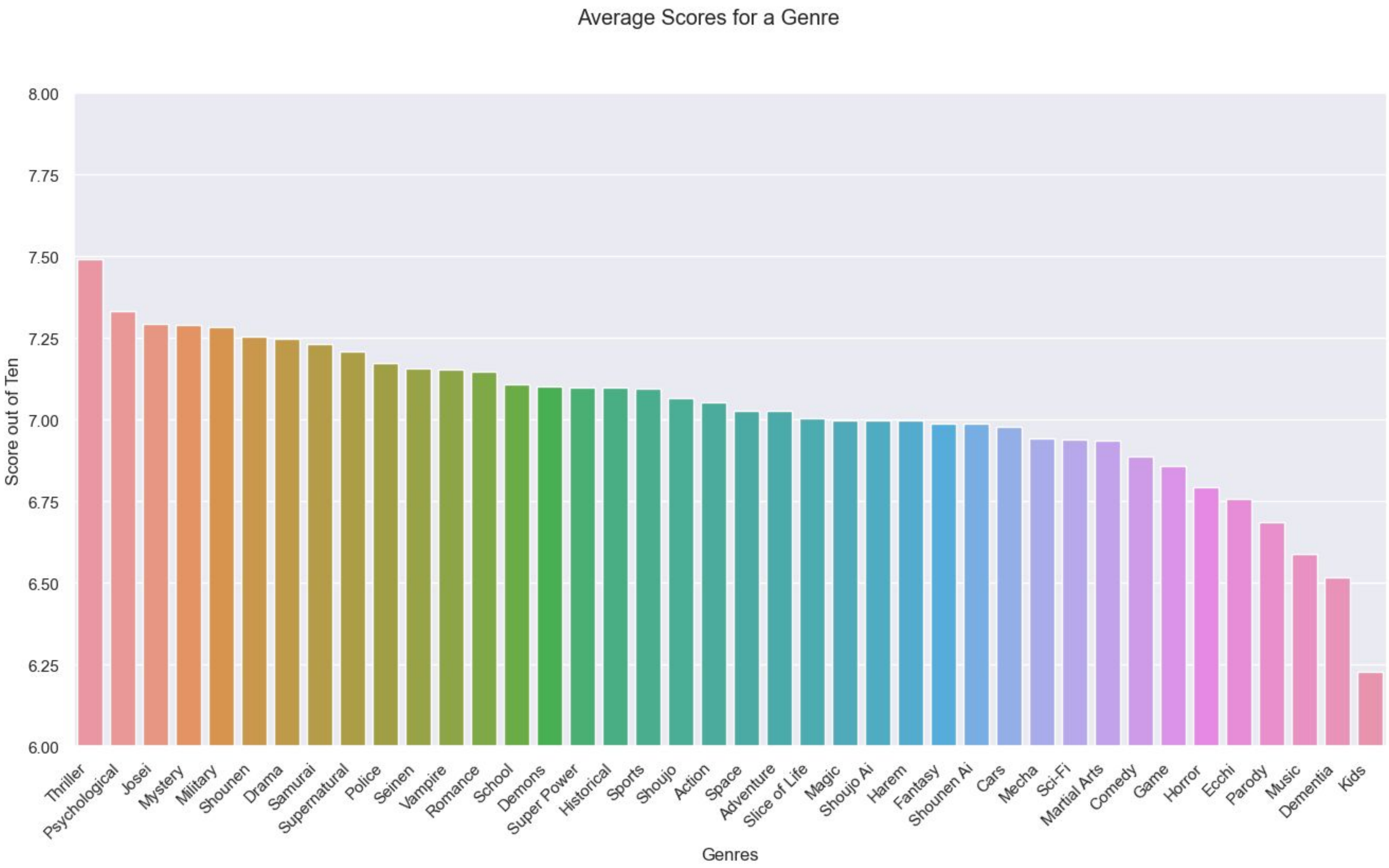

Figure 2.6: Average Scores for many Genres

Also, even after graphing the Top 10 Highly Rated Genres over the years (Figure 2.7), it seems that there is no genre dominance in terms of ratings, but there have been so interesting dips and beginnings. Military and Thriller animes seem to have had rough starts both starting out around the 5 range but eventually rose up to be one of the most well received genres. Another interesting dip was with Psychological where around 1985, the genre had a rough year dipping below 5. The scores might be this volatile because there are relatively little amounts of these genres out and the ones that are made might dictate that year’s scores.

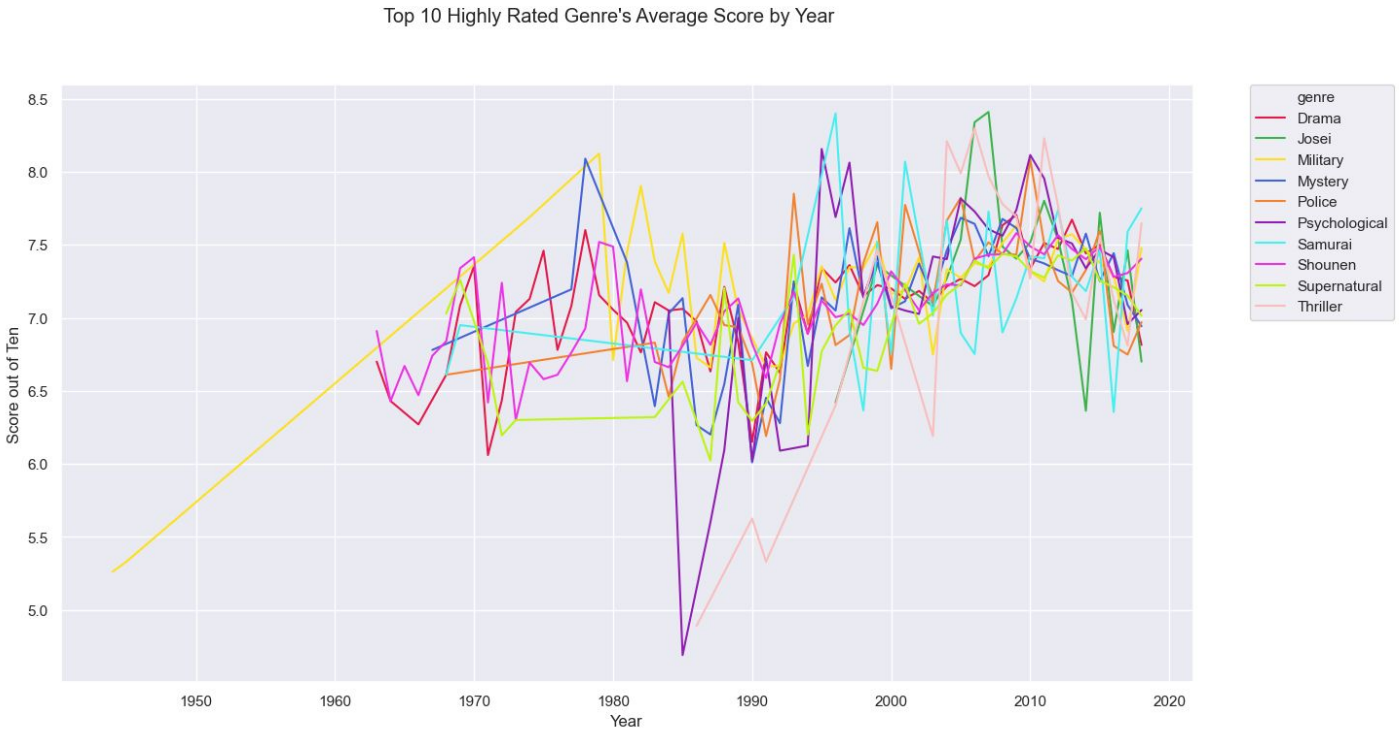

Figure 2.7: Average Yearly Scores for Top 20 Genres

I believe that this genre information can prove invaluable for licensors/producers. With statistical information on the most well received genres and a perspective on how saturated the market is with specific genres, they can optimize their anime for the best reception. For example, we can see Action is starting to overtake Comedy in saturation and this could mean it might be time to put a lot of effort into Action to ride along an existing trend. Also having information on top genres can help studios focus on aspects people tend to like, therefore making an anime with a wider, but at the same time, a more targeted appeal.

Research Question 3: What varies between different animation studios? Is there a "best" studio?

"The most highly rated studio is Studio Chizu, which is closely followed by Egg Firm. The worst rated studio was Three-D with Studio! Cucuri being the worst. With the most highly rated genre Thriller, White Fox was the most highly rated Studio and Production Reed was the worst."

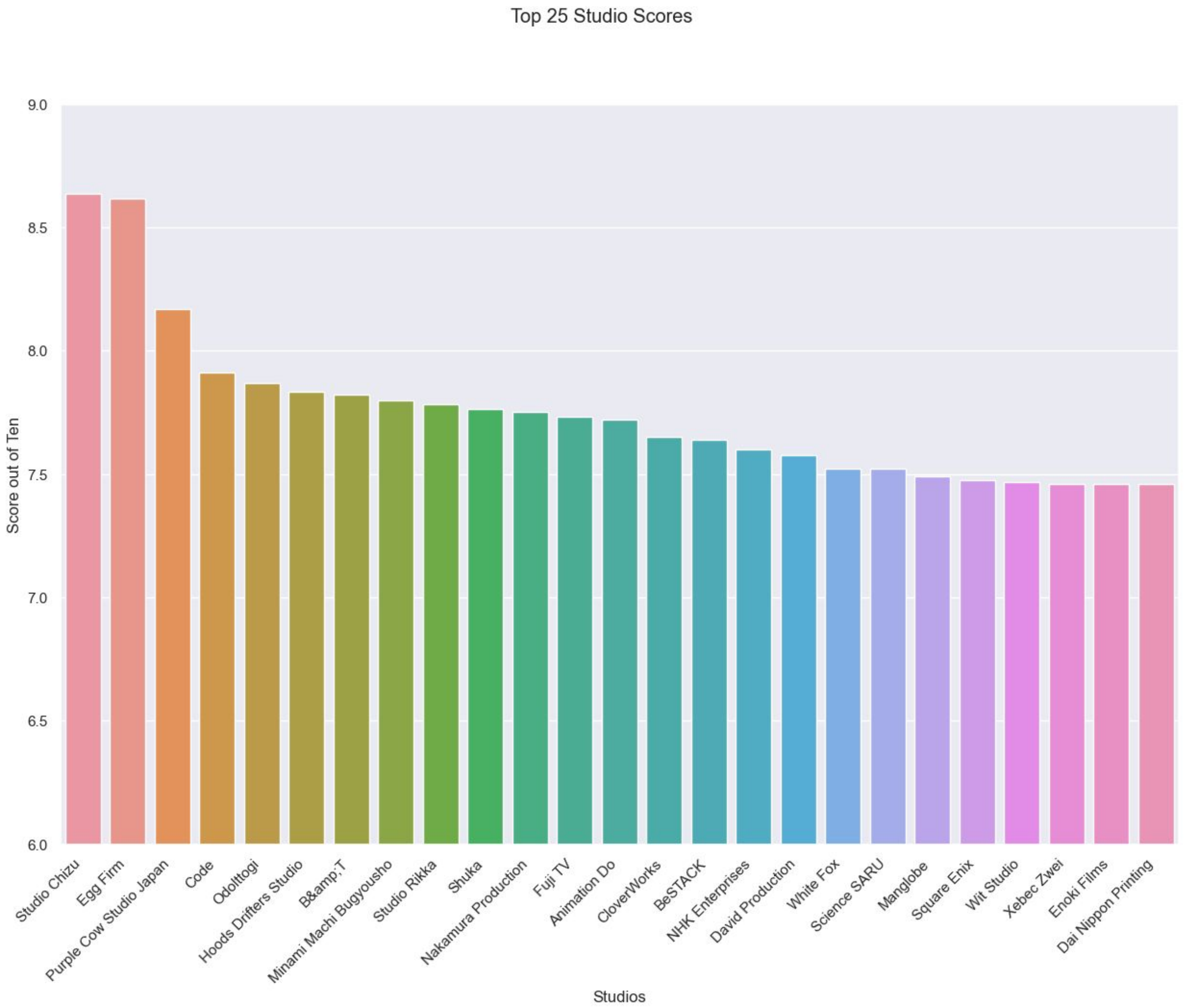

Figure 3.1: Average Scores of Top Studios

Taking a look at Figure 3.1, we can see Studio Chizu and Egg Firm are in the lead by a pretty significant margin compared to the rest of the studios. I was surprised that these studios were on top rather than acclaimed studios like Studio Ghibli or Shaft, both sporting series that are beloved by anime fans overseas and in Japan. What I noticed within these top 25 studios was that most of them have relatively small amounts of anime made, with Studio Chizu only having 3 (as of 2019). This indicates the similar principle of "Quality over Quantity" as bigger studios with more animations have their own ups and downs when producing so much. Smaller studios like Studio Chizu can make a few great animes and have their scores very high because they spend their time making sure one show is the best they can make. However, I believe this sacrifices popularity because these studios in the top 25 are not as popular as studios like Toei, A-1, or Kyoto Animations. When large studios make a lot of anime, they have a lot more audience that they can reach and get more popular, even though they might lose some average score due to not having every show be a mega hit.

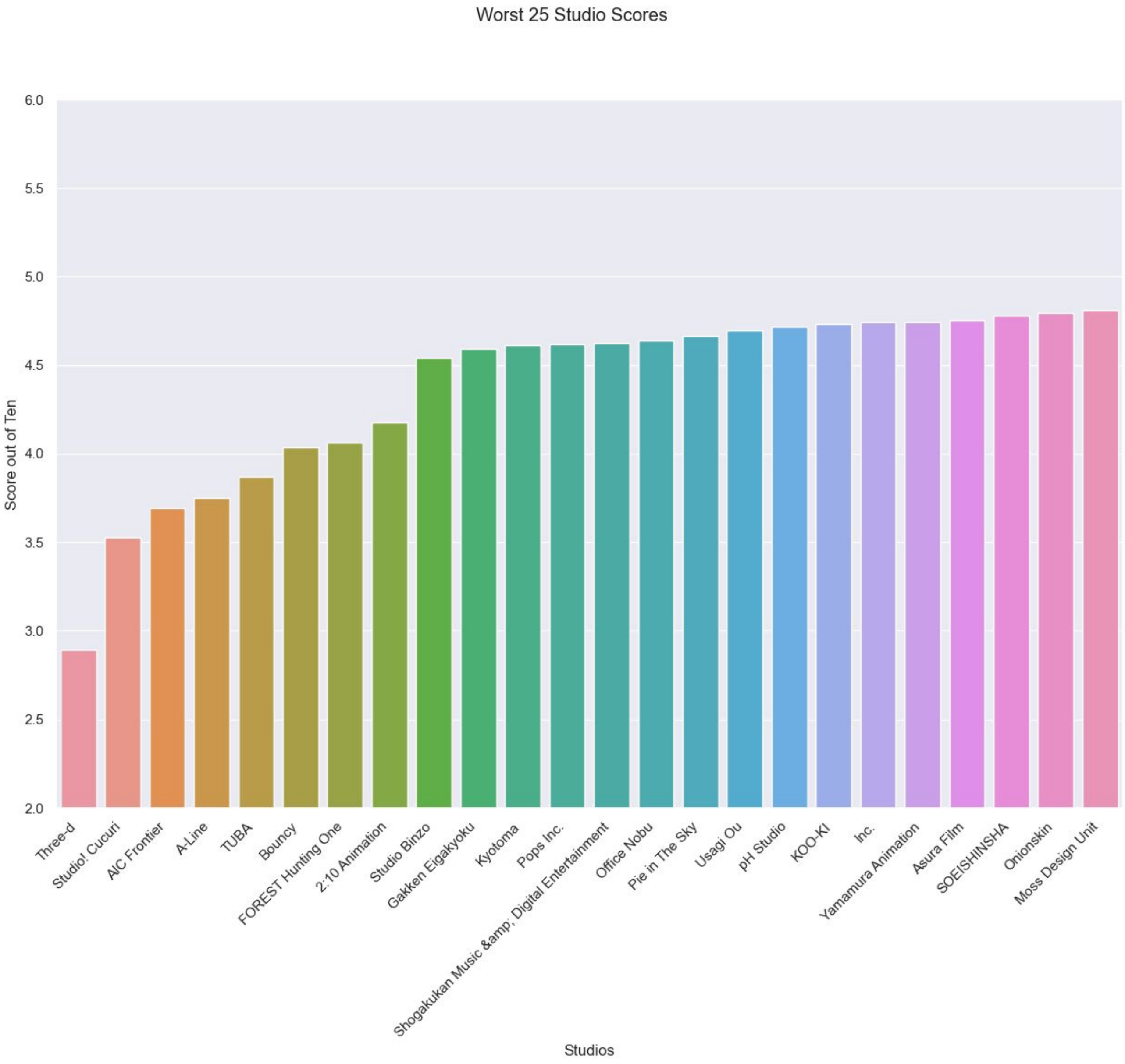

Figure 3.2: Average Scores of the Worst Studios

The pattern of smaller studios being at the extremes continues in Figure 3.2 as most of the studios in the bottom 25 also have less anime and they have been failures with the average scores being below 5 which is very low considering the average is as high as 8 for all anime.

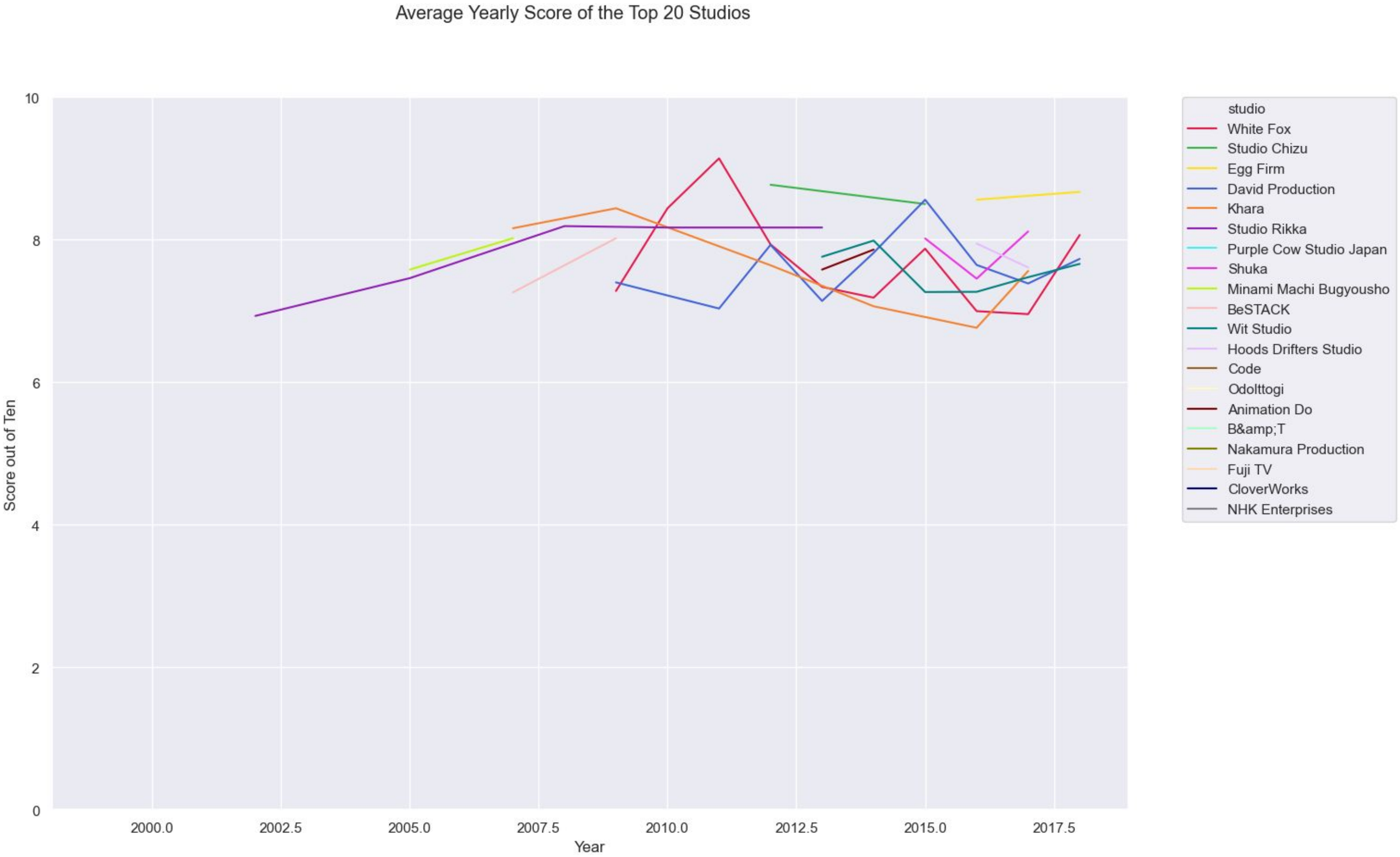

Figure 3.3: Average Yearly Scores for the Top 20 Studios

When graphing the top 50 studio's yearly scores (Figure 3.3), I found that there hasn't really been "studio dominance" within the anime industry for highly rated shows. All the studios seem to range from 7-8 and none had a "rough start." Even when extending to the top 50 studios (Figure 3.4), the range seems pretty contained, indicating that there isn't really a "best" studio that stands atop the others in a significant way.

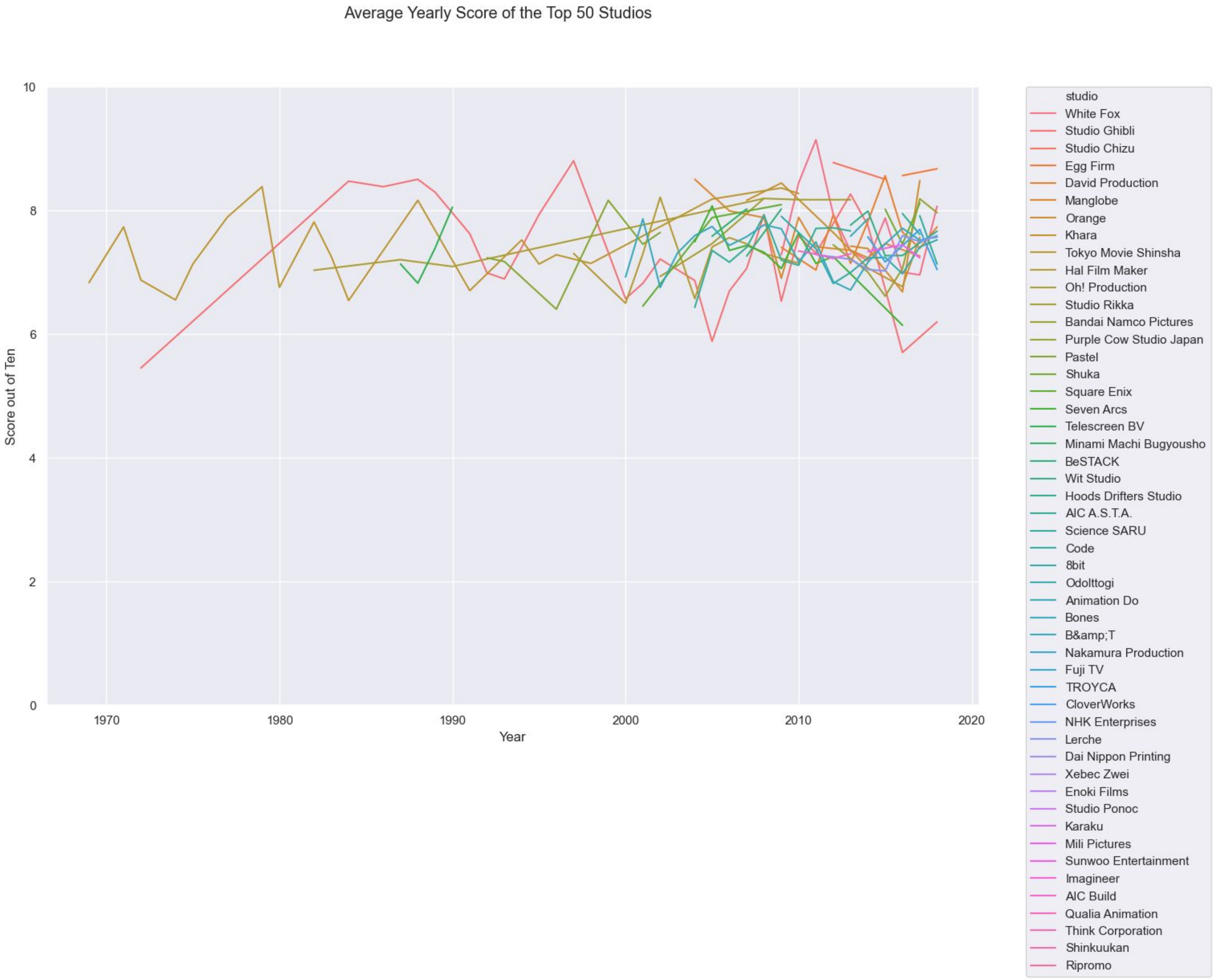

Figure 3.4: Average Yearly Scores for the Top 50 Studios

However, when looking at how much anime a studio has made (Figure 3.5), it can be seen that Toei Animation has been “dominating” in terms of anime output. Being at over 400 animes compared to Sunrise’s ~300, Toei Animation leads the anime industry with efficiency and work output. This might be expected though as they are responsible for some of the biggest anime franchises like One Piece or Dragonball. However, if we look at the yearly anime creation rates (Figure 3.6), we see that Toei’s output has fallen in line with the other popular studios, indicating slow growth for the studio. A-1 Pictures seems to have taken the lead in modern shows with the most anime output after 2010, especially when they have made hyper popular shows like Sword Art Online and Fairy Tail.

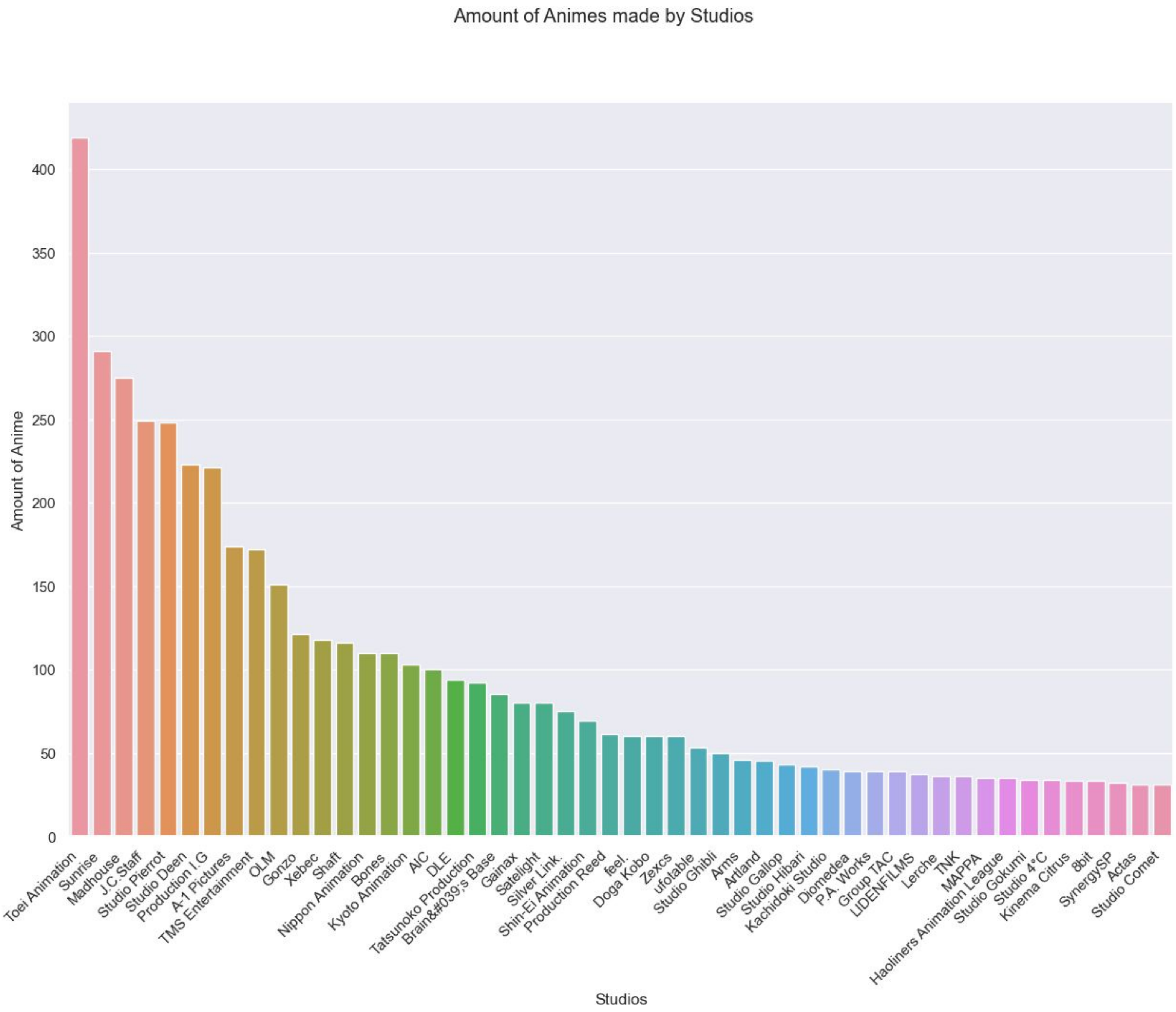

Figure 3.5: Studio Anime Counts

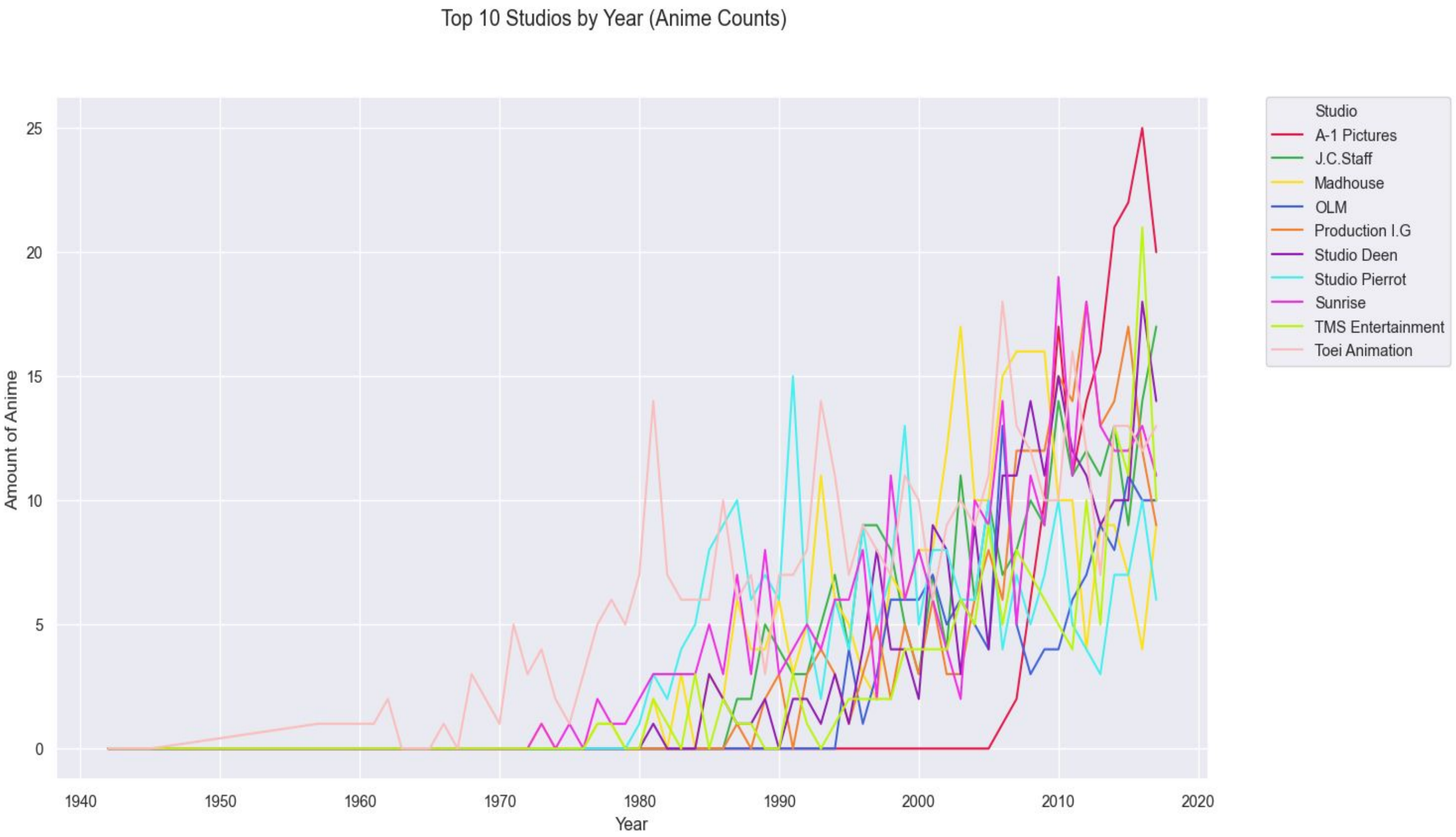

Figure 3.6: Studio Counts Per Year

Figures 3.7-3.12 are all bar plots that show the best and worst studios for certain genres. For anime producers/licensors, this information could prove very useful as it can help ensure a specific anime they want to make is made by a studio that knows how to make it (or at least by a studio who doesn't have a bad track record.) If you want to see more genres, please take a look at the Google Drive link at the beginning of the Results section. It will be in the folder, "rq3_genres".

With information on what studios specialize on, anime producers and licensors have a huge advantage when it comes to selecting a studio that can be a good investment. Licensors can also see which smaller studios churn out the best scoring anime and stay away from the ones that have consistently failed. This information can prove helpful for the increased output of anime that are created with higher quality and by professionals that are attuned to the genre.

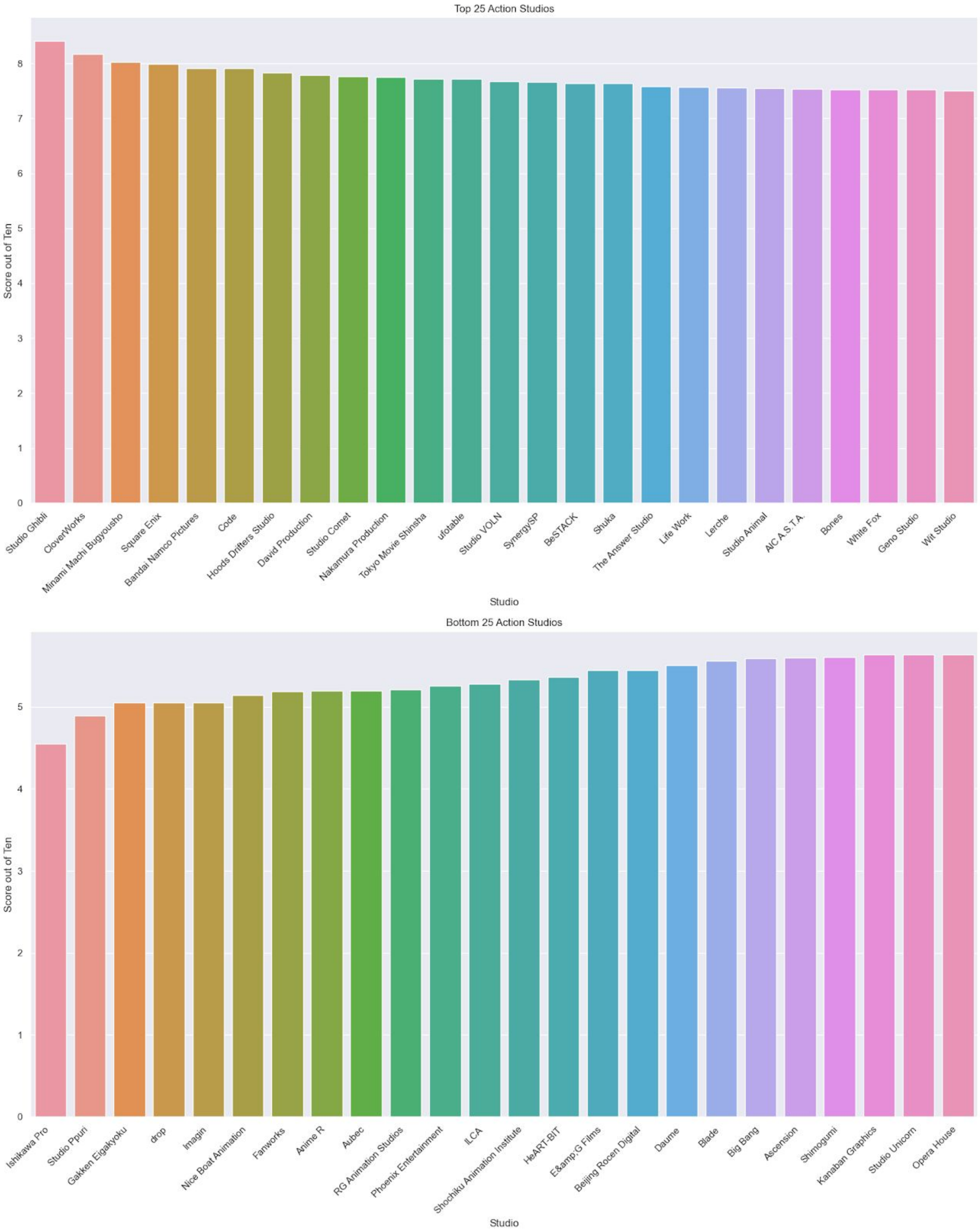

Figure 3.7: Best and Worst 25 Action Studios

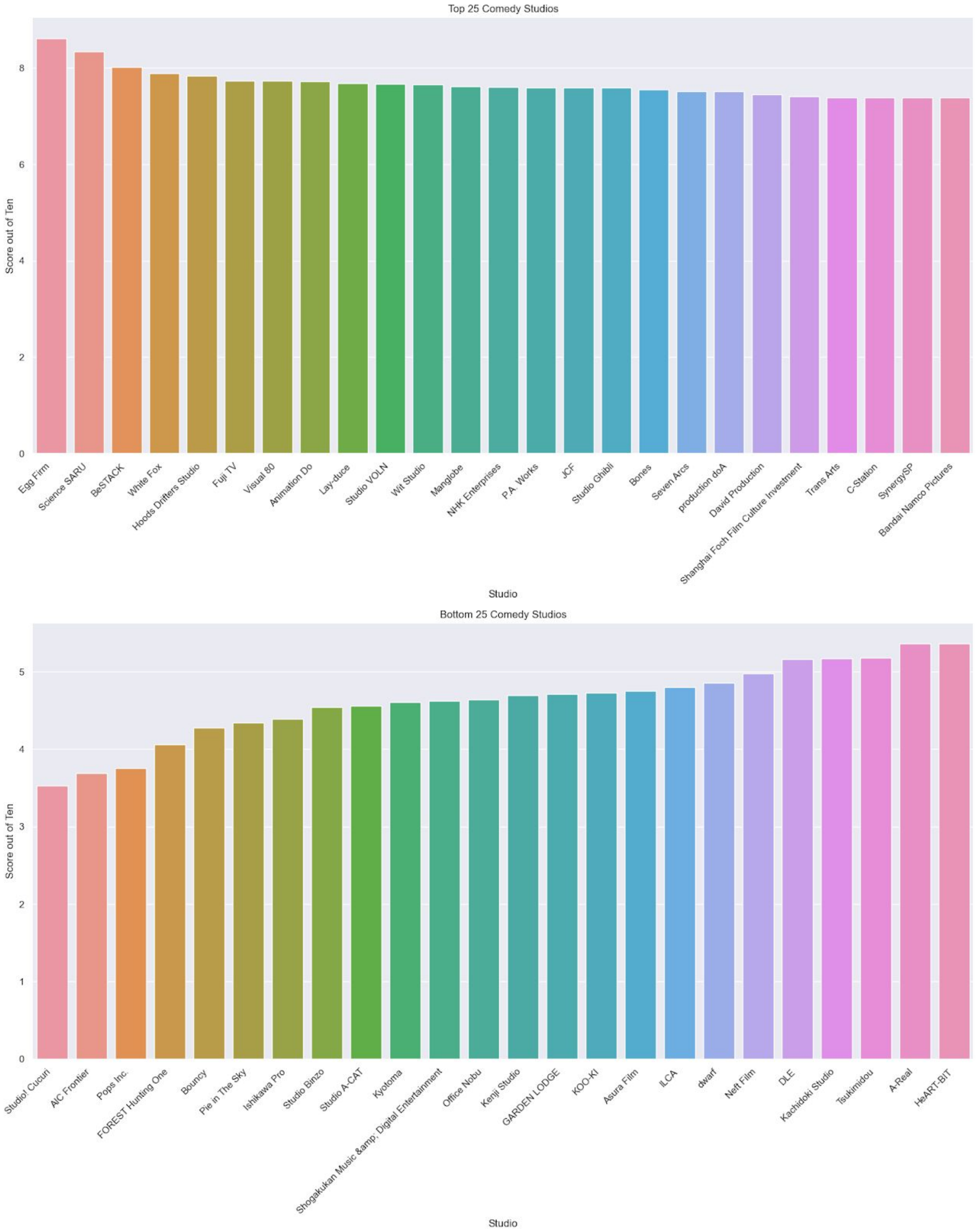

Figure 3.8: Best and Worst 25 Comedy Studios

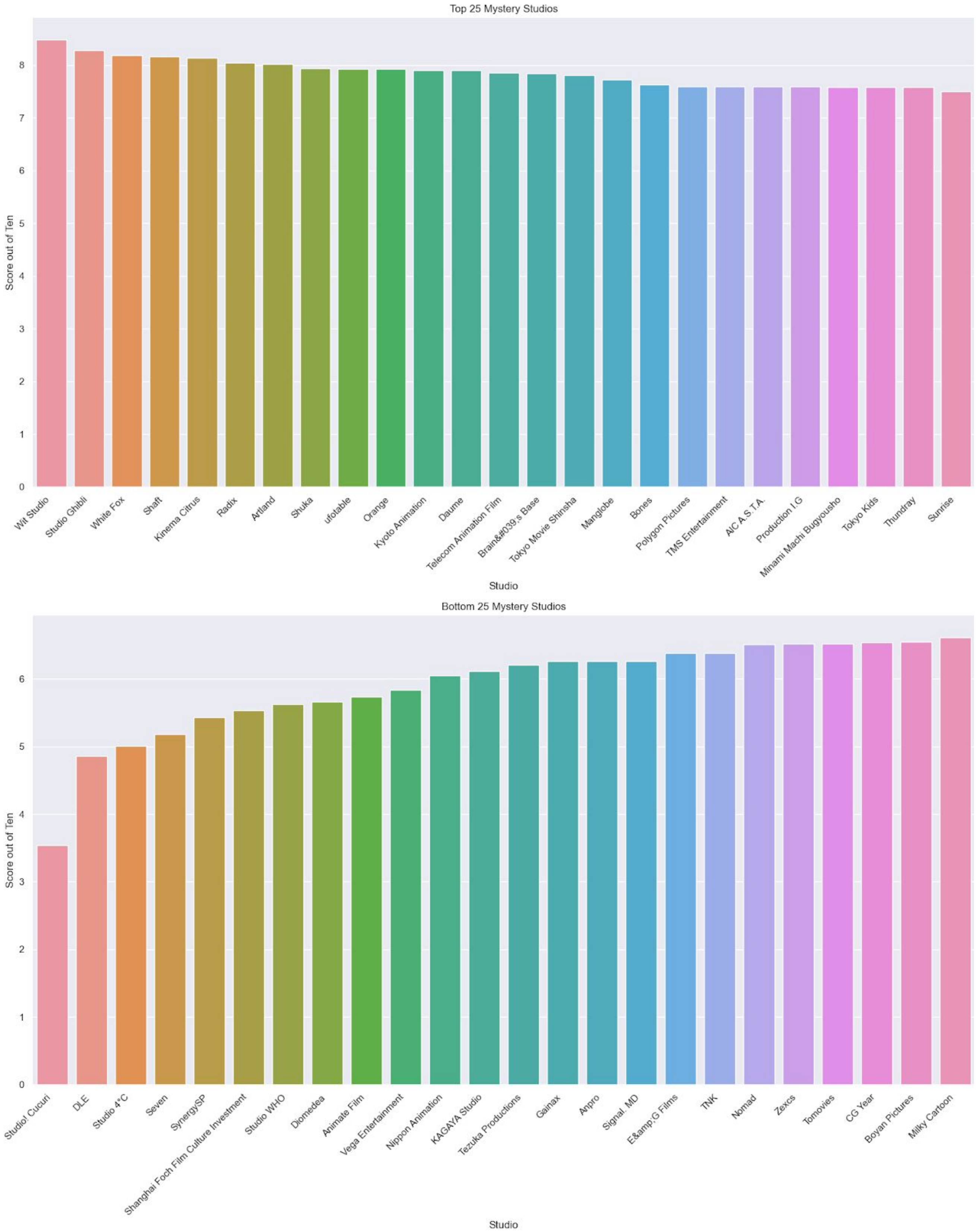

Figure 3.9: Best and Worst 25 Mystery Studios

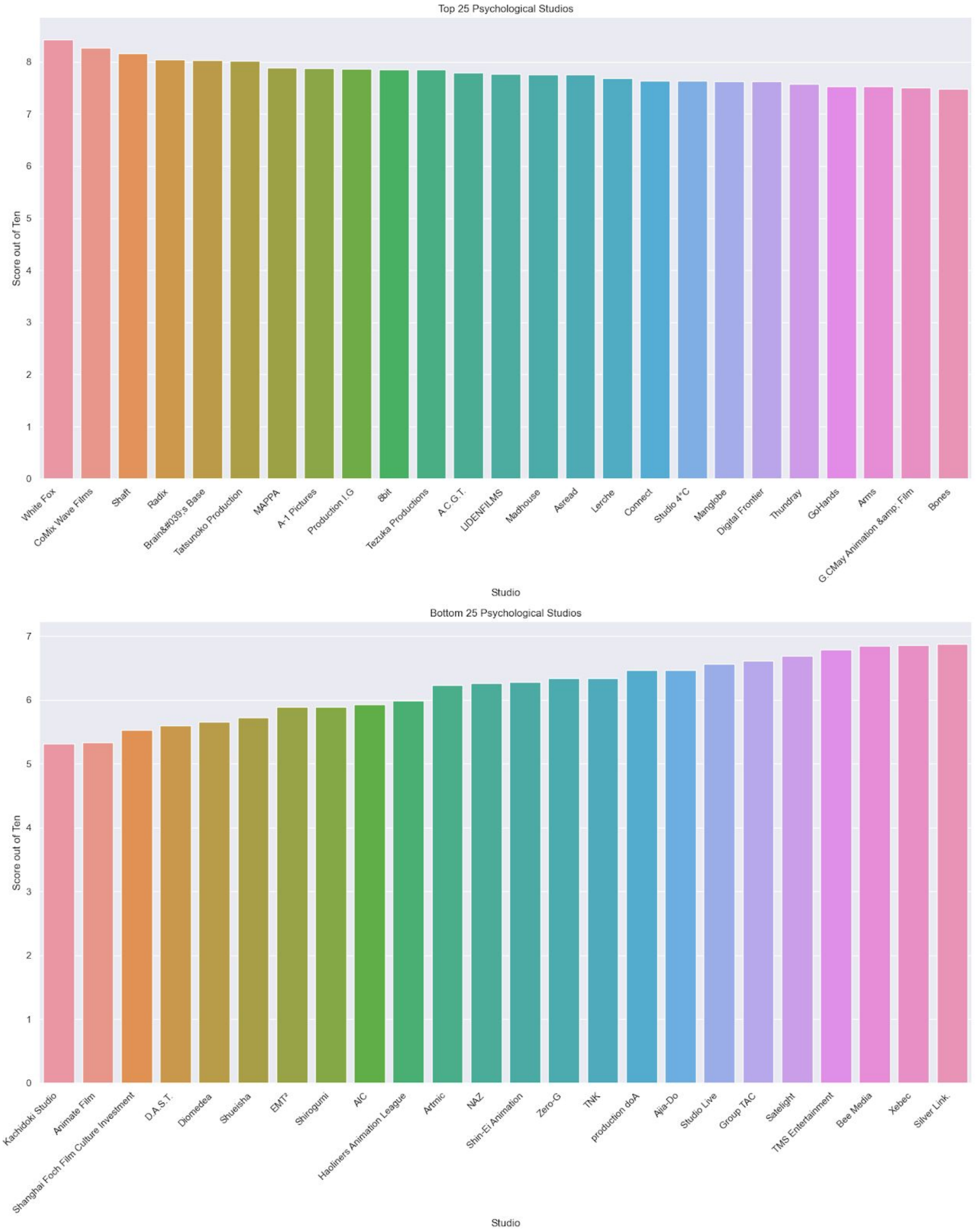

Figure 3.10: Best and Worst 25 Psychological Studios

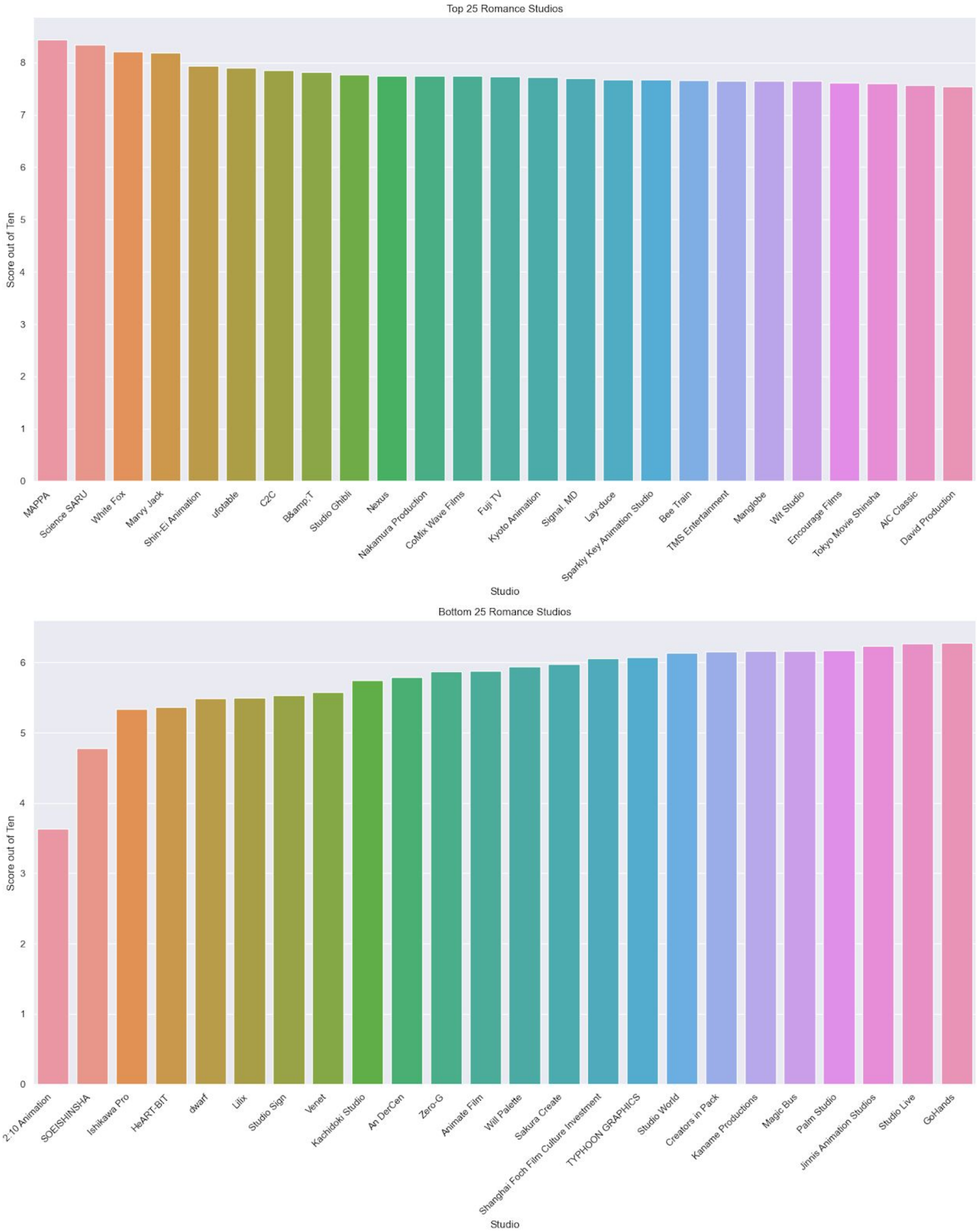

Figure 3.11: Best and Worst 25 Romance Studios

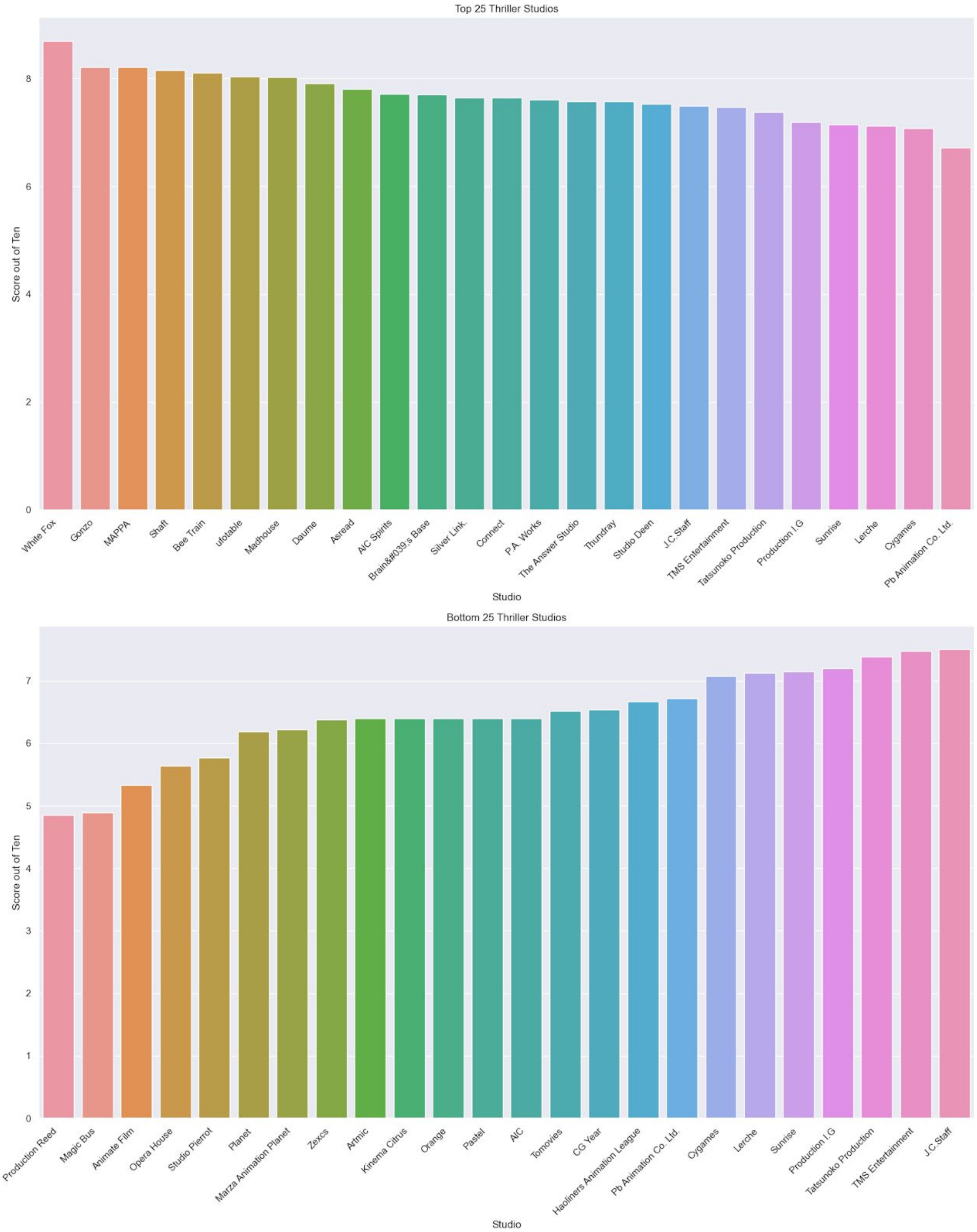

Figure 3.12: Best and Worst 25 Thriller Studios

Research Question 4: Can we predict the score of an anime on MyAnimeList using factors like genres, animation studio, air date, source material, and etc? What do we change to maximize a score?

"The machine learning models were fairly successful with the Mean Absolute Error for score being less than one point off and the favorites being around 800 people off. It seems the most important factors to maximize an anime's score is its duration and studio. For popularity (favorites), it seems the two most important factors are the type of anime (TV, Movie, Special, etc) and the amount of episodes it has."

When making this machine learning model, I decided to go with a Decision Tree Regressor Model because it allowed me to predict numerical values and it was a model that was appropriate for my current skill level with machine learning. After training my first model without any hyperparameters (the depth could go as deep as it wants) I got around 73 levels for the tree. The Mean Absolute Error for that model was around 1 point off. However, to determine the best hyperparameters I decided to train a model for each hyper parameter from 1 to 49 in intervals of two (Figure 4.1). What I found out is that for the score model, the best hyper parameters were around 7-11, giving around 0.65 points off with Mean Absolute Error. When going beyond 15, it seems the score error became worse and worse. Mean Squared Error mirrored similar results but is a bit more extreme as it squares the errors.

For the popularity model, the error was much larger with errors being around 800 people. However, with popularity that can reach the millions, 800 is quite small. While not specifically accurate, the model can generally help predict a certain anime's general popularity. However, what I noticed is that changing the hyperparameter seemed very random when impacting error. This indicates to me that the popularity model is not as trustworthy as the score model and there are most likely a lot more factors that impact popularity. If I had to take a guess, I would think factors like social and cultural context would be important for popularity, even though those two are outside an anime studio's reach when creating an anime.

To determine what aspects of an anime impact score and popularity the most, I removed features one by one and took note of the error change (Figure 3.2). If the error went up, I assumed that the feature was important, and if the error went down, that feature might be negatively impacting the data or having little effect. For score, the most important factor seemed to be duration and studio. The other features seemed to have less impact and I decided to disregard the Mean Squared Error change because it seemed to inflate the errors while Mean Absolute Change seemed more accurate. For popularity, it seems that the most significant factor would be the anime type that it is, which ranges from TV shows, to movies. I was surprised by the scores' dependence on duration because I thought that would be an indifferent feature. For popularity, I was also surprised because I thought the studio would determine popularity the most.

Having a model that can predict anime scores fairly accurately and have a general idea of how popular an anime can be, is incredibly powerful, and allows for less money to be wasted on endeavours that might have failed.

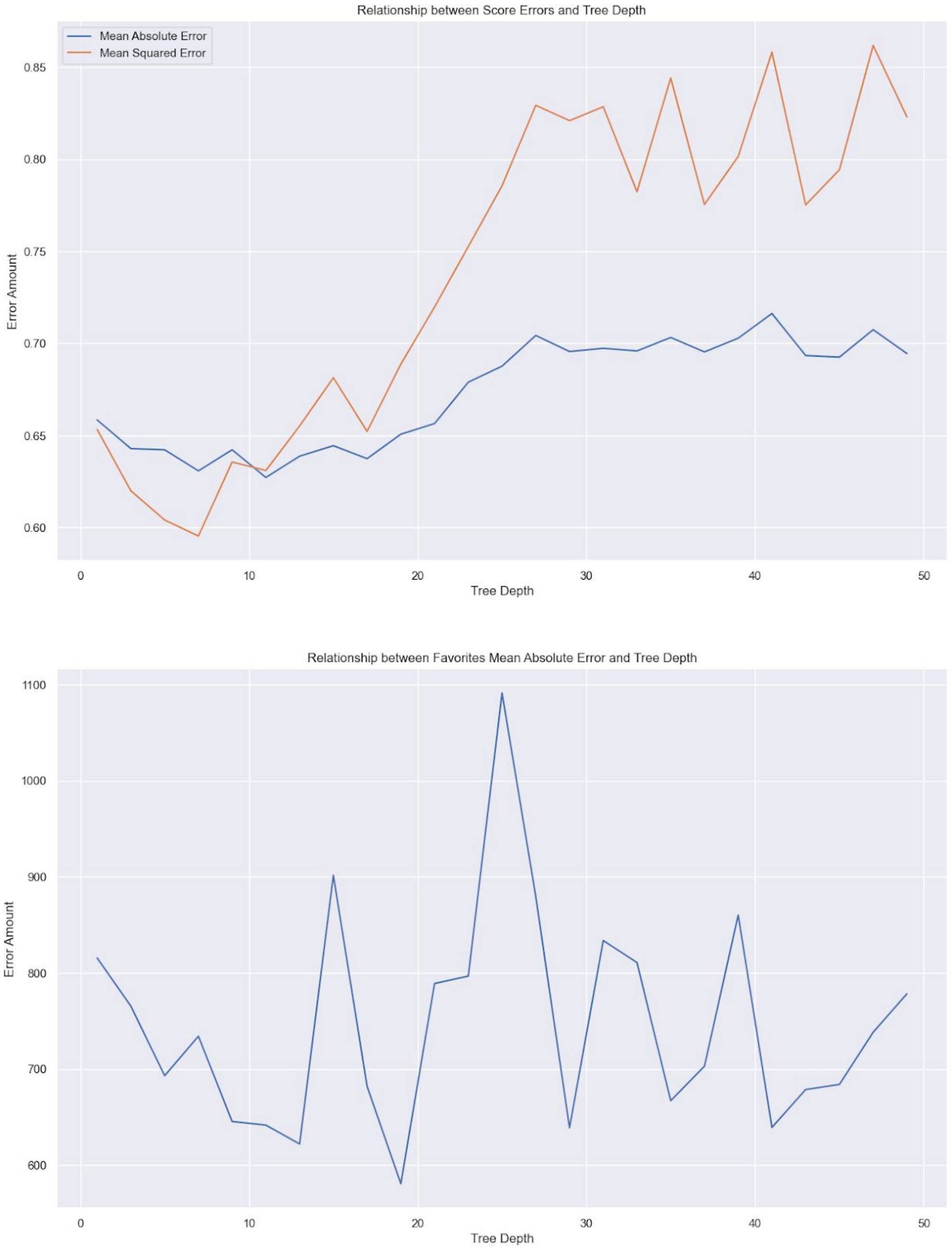

Figure 4.1: Line Plot of Errors Based on Tree Depth

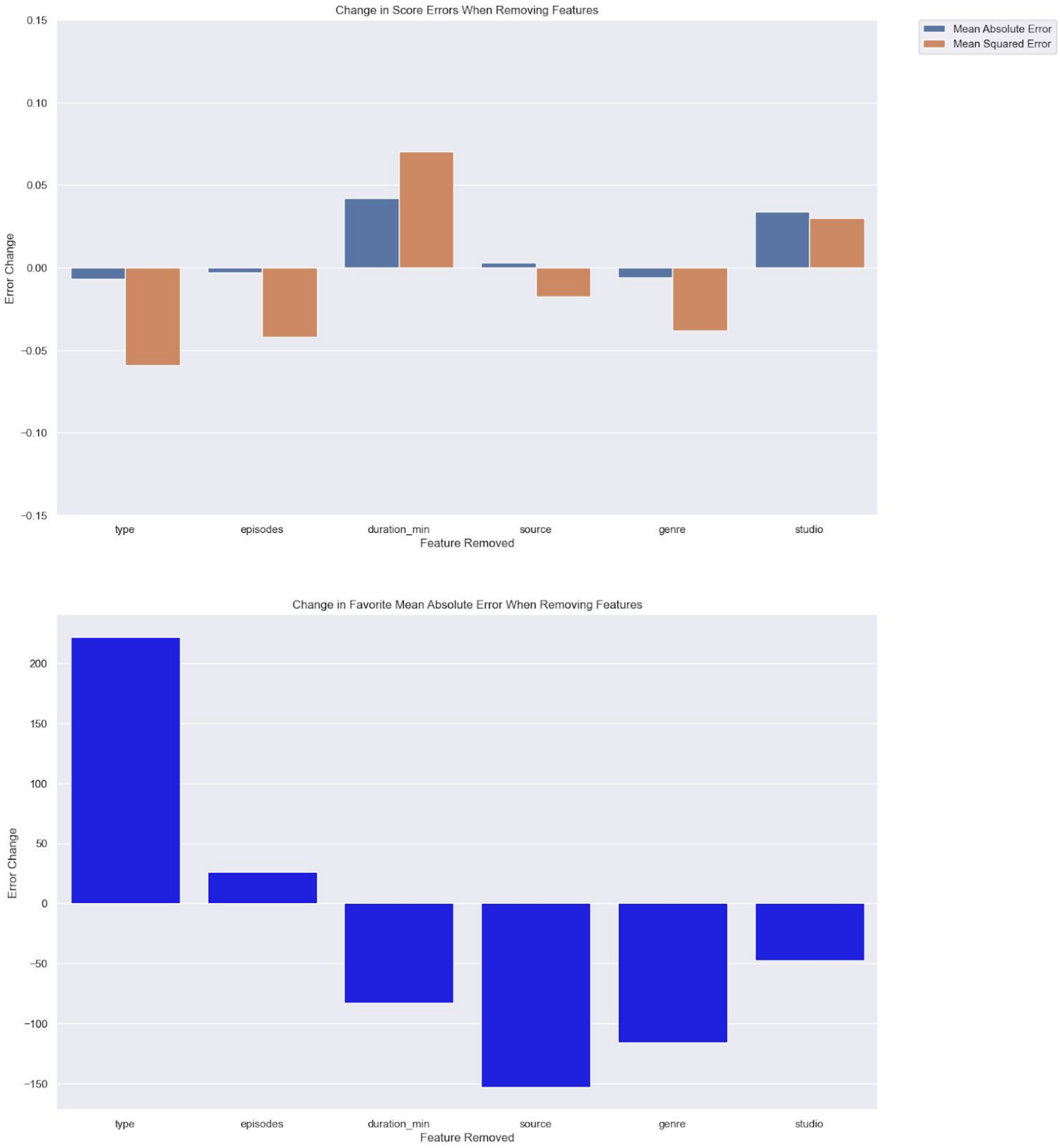

Figure 4.2: Bar graphs of Error Change

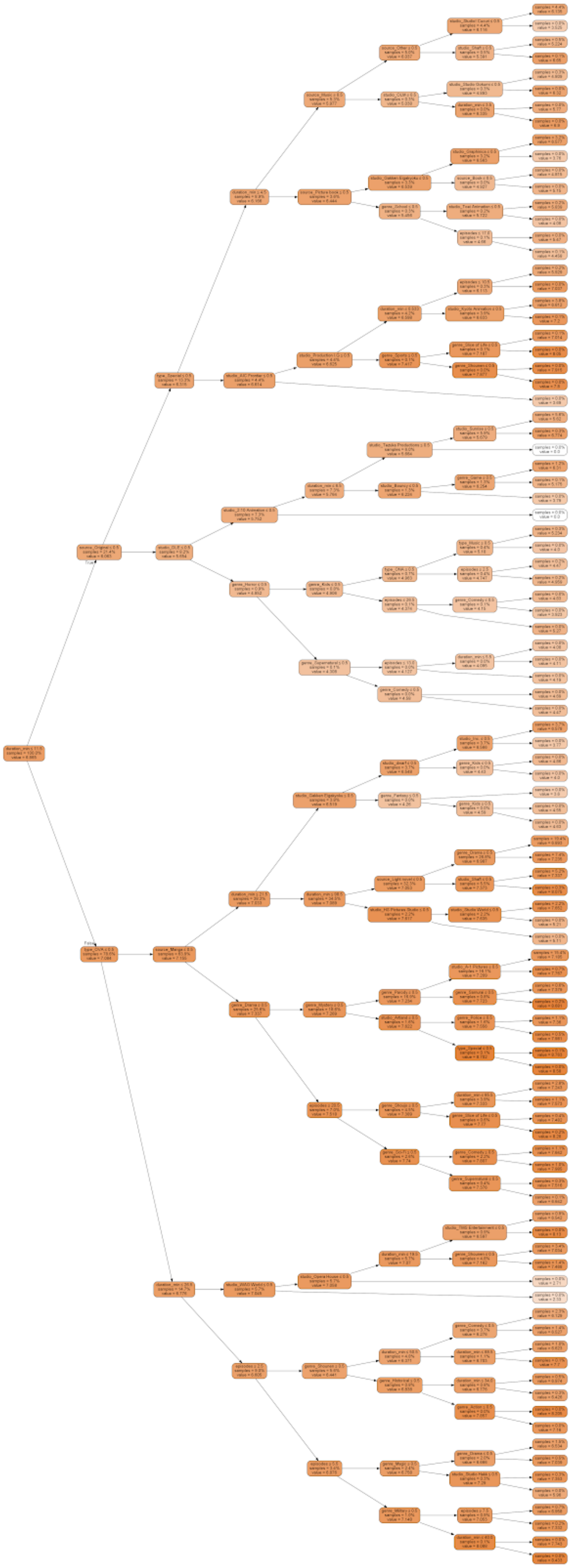

Figure 4.3: Decision Tree for Score Model

Challenge Goals:

Multiple Datasets:

Within the MAL data, there are three unique datasets that can be used to create a richer analysis of anime and anime culture in general. The three are information on animes, the anime users, and the animelists that each user has. Each of these datasets have little overlap and when they do, it is usually one or two columns to effectively connect the datasets in a meaningful way. For example, the user animelists have a username that can be found in the user list and they have anime IDs that can be used to tie it to the anime list. I believed that these datasets within the MAL data were used in tandem to warrant the challenge of combining multiple datasets to make a richer analysis.

Machine Learning:

For Machine Learning, I created the most accurate model I can to predict the scores and popularity of an anime series given specific aspects. I used the model to create an "optimal" anime and potentially learn new and interesting facts that might not be possible to glance from traditional data analysis. The challenge for this approach is that I did more research into machine learning, determining the best model for the approach and adjusted specific hyperparameters to optimize the model. I also looked into entirely different ways of thinking of making the model such as "top-down" or "bottom-up" approaches.

Messy Data:

To provide more grounded and valid tests for my machine learning model, I created my own dataset 2019 anime data to see if my machine learning model is accurate at all. To get this data, I used the unofficial REST API, Jikan, and an accompanying Python wrapper to retrieve MAL data. I think this meets the needs of the "Messy data" challenge because I used an external API to gather the data and pre-process it to fit in with my calculations.

Work Plan and Evaluation:

Work Plan:

Data Retrieval/Cleaning (May 23) approx. 3 hours
RQ 1 (May 24 - 25) approx. 4 hours
RQ 2 (May 26 - 28) approx. 5 hours
RQ 3 (May 29 - 30) approx. 4 hours
RQ 4 (May 31 - Jun 3) approx. 7 hours
Gathering all visualizations and writing a report (Jun 4 - 5) approx. 4 hours
Creating Presentation (Jun 4 - Jun 11) approx. 4 hours

**See above for specific plan for each section*

Evaluation:
Overall, I believe I underestimated this project and the time I would put into it as a whole. I also was pretty off on all of my approximate time estimations for this project.

For Data Retrieval and Cleaning, I believed it would be one of the most simple parts of the project. For cleaning, that remained true as all I had to do was filter out columns I would not use in my data and create new CSV files to use. However, when it came to getting 2019 anime data, it was a lot harder than I thought. I knew data scraping would be challenging but the biggest problem I faced was with the actual API rate limiting rather than the code. Jikan limits every user to 1 request every 4 seconds and I had to make 2 requests for every anime in 2019. This made it harder to test and made the scraping process last a lot longer than I would have liked, pushing my finish date to around May 24.

Since everything was delayed, I decided to shorten one of my questions and that came down to the Research Question 3. For the rest of the research questions, everything went pretty smoothly but took 25% more time to finish than I initially predicted. The biggest contributor to the slowdown would be me figuring out how to separate multi-studio and multi-genre anime while still retaining proper information. After figuring that out I was able to go through the rest of the project.

One thing that totally went over my head was to write proper documentation and do testing. I completely forgot that I had to do these two things and didn't factor it into my Work Plan. This cost around 4 hours on top of the extended hours of the other days. Because of the complications, I ended up spending more time each day than I originally expected.

Properly organizing and detailing my research was also time consuming with image formatting and writing explanations for every one of my observations. This similarly cost me ~25% more time to create the report.

Despite these complications, I was able to make my other deadlines within my work schedule, and believed this project overall was a fun way to dive deeper into a medium I have recently begun to enjoy.

Testing

For me, testing was a difficult thing to do. I couldn't really think of effective ways to test my data because I had already finished up all of my research questions before realizing I needed to do testing. All of my code was rigorously debugged with countless print statements to manually see if all my data manipulation was correct and I made sure the graphs I made were sane. I generally used “common sense” with my code and tried verifying if it looked right when initially making the code. But that doesn’t leave behind proof because print statements are temporary and are only for the testing phases.

However, I did try to implement proper tests afterwards which are included in tests.py. I didn’t do a unit test for each function, but I did use a smaller subset of data to manually verify if common calculations I did were proper and produced the desired result. I usually used the built in assert in Python to help me with verification.

For my machine learning research (RQ4), I combined my testing and actual research problem with how I wanted to verify my machine learning results through graphing its error. I graphed error in relation to hyper parameters and features. Because the popularity error was a lot more erratic, I decided to focus less on that and focus more on score. Graphing allowed me to “test” the validity of the machine learning models to a certain extent.

Collaboration

This was a project that was completed by KV Le (Me). I mainly used two online resources which were Online Documentation for my included libraries and StackOverflow.