CSE46801 - Information Visualization
Assignment 1: Anime Data Analysis

Student name: Nguyen Minh Duc
Student ID: 20202026

1 Introduction

The anime industry has been growing rapidly in recent years. More and more people have started watching this particular style of animation from Japan. Understanding the preferences of anime fans would benefit studios in production and marketing. Moreover, as the industry is growing, more and more anime series and movies are produced creating a large database, which could be utilized to understand the most popular genres, themes, and characters, helping the studios to create content that is likely to resonate with their target audience. Lastly, analyzing anime data can also aid in the study of Japanese culture and its influence on popular media worldwide. By analyzing trends and patterns in anime consumption and production, one could gain a better understanding of the cultural and social factors that shape media consumption and its impact on society.

In this assignment, I will use "Anime DataSet 2022" from vishalmane10 [1] containing 18495 entries of different animes. The data is crawled from "Anime Planet" [2], a well-known database that stores new anime series, movies, and manga (Japanese-style comics) every season. In this paper, I will discover three main questions:

- What anime has the most ratings of all time?
- What is the trend of anime type?
- What is the most popular anime genre?
- What is the best anime studio?

and some follow-up questions. To do the analysis, the following open-source Python libraries are utilized: numpy [3], pandas[4], matplotlib [5], and plotly [6].

2 Data Preprocessing

The data set contains 18495 rows and 19 columns of different attributes for each anime, namely, Rank, Name, Japanese name, Type, Episodes, Studio, Release season, Tags, Rating, Release year, End year, Description, Content Warning, Related Manga, Related anime, Voice actors, staff. The data set contains a lot of `NaN` values and most of them are unstructured. After consideration, I decided to keep some important attributes that will be useful for analysis: Title, Type, Studio, Tags, Rating, Release year, and Content Warning. All of the `NaN` values will be converted to 0 in the `rating` attribute, and remove every entry that has a `NaN` studio and tag, reducing the number of anime in the data set down to 11930 entries.

For this section, I utilized powerful `pandas`'s preprocessing methods that allow me to filter out unnecessary data, and to fill in some of the missing values.

3 The best anime of all time

One of the first intuitive things to do when we have this kind of data set is to see what is the most interesting movie/anime to watch. In this section, I will investigate what anime is the best, i.e. having the most rating according to the database, of all time. Since we already have the rating attribute in the data set, the job becomes trivial. I just need to sort the data frame in decreasing order of `Rating` with the following lines

```
sorted_df = df.sort_values(by="Rating", ascending=False)
top_10_rated = sorted_df.head(10)
```

where `df` and `top_10_rated` is the data frame containing the cleansed entries and the top 10 highest rating anime of all time. Finally, I just need to plot it out and obtain the visualization. Please refer to Figure 1 for more details.

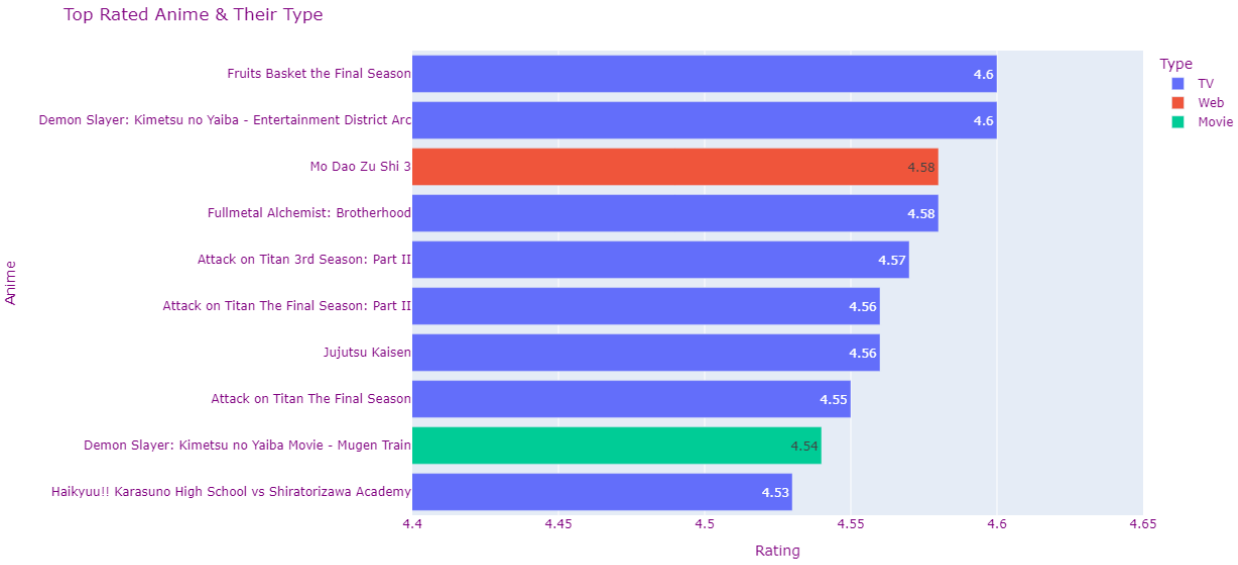

Figure 1: Top 20 most rated anime

One could immediately observe that there are two anime sharing the top spots: "Fruits Basket the Final Season" and "Demon Slayer", having the same 4.6 ratings. One interesting thing that could be pointed out from this figure is that eight out of ten are TV-series type of anime. This shows that the majority of the audience enjoys watching TV series rather than just one movie, or it could be the case that fewer anime movies are produced compared to TV series. Another interesting fact is that "Mo Dao Zu Shi 3", a Chinese animation based on a Web novel with Japanese-style art is sitting as the third most-rated anime of all time. Therefore, this data set does not just contain animations produced in Japan but also has animations from other countries with similar art styles to Japan. However, relying on the ratings alone does not tell objectively which anime is the best to watch. The average rating of a show also depends on the number of votes it receives. To further investigate what is the true "best" anime to watch, we need to obtain the number of votes from users for each show. Unfortunately, the data set does not contain any information regarding this issue, so I have to crawl this data by myself using two Python open-source libraries: `requests` [7] and `BeautifulSoup` [8]. Inserting this newly crawled attribute into the data frame, I can obtain a scatter plot for the top 100 rated shown in the database. Please refer to Figure 2 for more information.

We can observe from the figure that the number of votes for web-type anime is fairly low, and most of the dots are blue, i.e., of TV-series type. There are two outliers: "Attack on Titan" on

Figure 2: Top 100 most rated anime

the top left, and "Fullmetal Alchemist: Brotherhood" on the top right. Both of them are very popular among anime fans. The majority of the show concentrates on the lower left region of the grid, which makes sense since not many shows have both high ratings and fans to vote for. This distribution of votes and ratings prompts another interesting question: What is the overall rating distribution of the database? By aggregating the ratings into small bins, I can easily obtain the distribution, please refer to Figure 3 for the visualization. One could easily observe the normal distribution of data with an average rating of 3.355, which is expected from a large sample such as this data set.

4 The trend in anime type

The second thing I want to investigate is the trend when producing different types of anime, i.e., is TV-series, movies, music, etc. To see the trend, I need to count the anime types for each year and visualize them like a time series. Doing this requires some preprocessing, the anime types in the database have unnecessary white spaces which need to be removed, which is done by using the standard `strip` method in Python.

```
for i in range(len(df["Type"])):
    df["Type"][i] = df["Type"][i].strip()
df["Type"].unique()
```

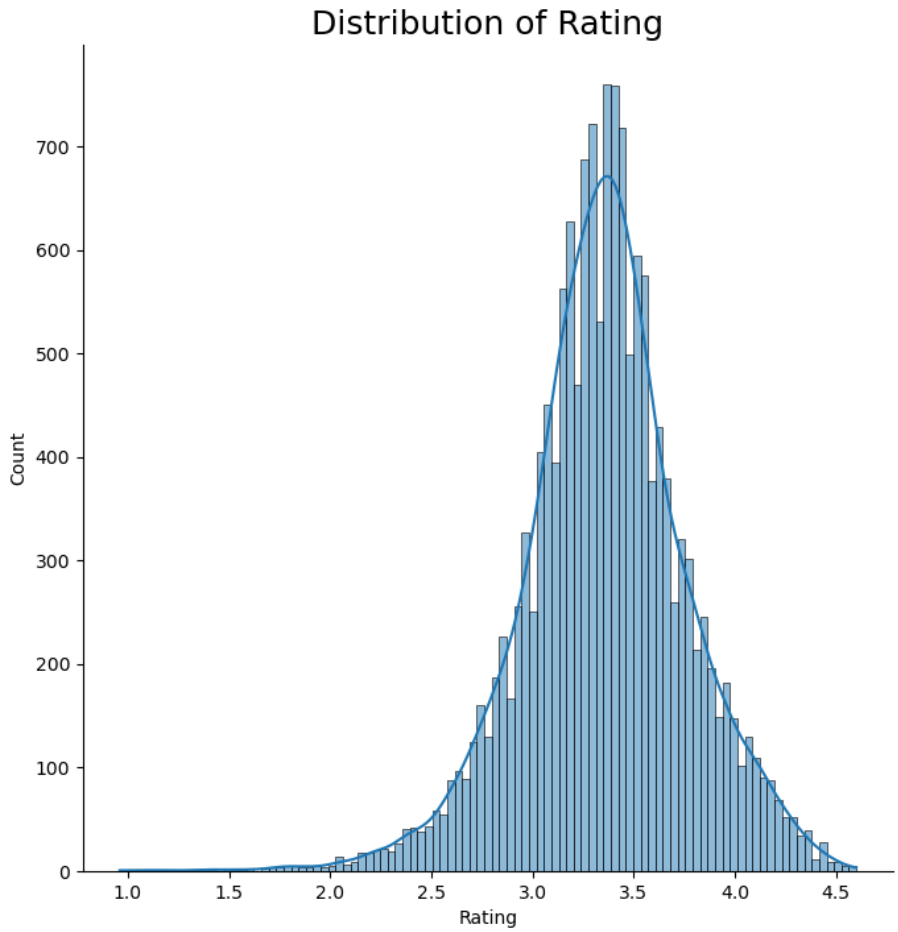

Figure 3: Ratings distribution

This results in 8 different types of anime:

```
array(['TV', 'Web', 'Movie', 'OVA', 'Music', 'TV Sp', 'DVD S', 'Other'],dtype=object)
```

After this, I need to group all of the data with respect to `type` and `release_year`. The result is Figure 4. As one could observe from the figure, most of the anime ever produced concentrated in the past 20 to 30 years of the industry's history as this is when colored TV started to get its popularity in the world because of its cheaper price and convenience. This is also when anime starts to spread widely across Japan, and even "exports" this kind of culture worldwide. Another interesting fact is that around 2010, web-based anime skyrocketed in a short amount of time thanks to the boom of the internet, more and more web-based animations are produced every year, and the number of web-based anime is now comparable to the traditional TV series. Together, web-based and TV series anime are responsible for almost half of the produced anime in recent years. Another interesting rise is the Music anime, which gain popularity in the latter haft of the last decade, this could be explained by a cultural background where virtual singers are a very popular concept in Japan. Lastly, there is a counter-intuitive behavior in the trend is a sudden rise in DVD production in the last 10 years since DVD disks start to become obsolete in this digital era. This is because of the surge in the demand for high-resolution anime during the time when transmission of high-quality videos is still rare and expensive. The studio could sell physical copies of the anime that might contain deleted scenes when it aired on TV.

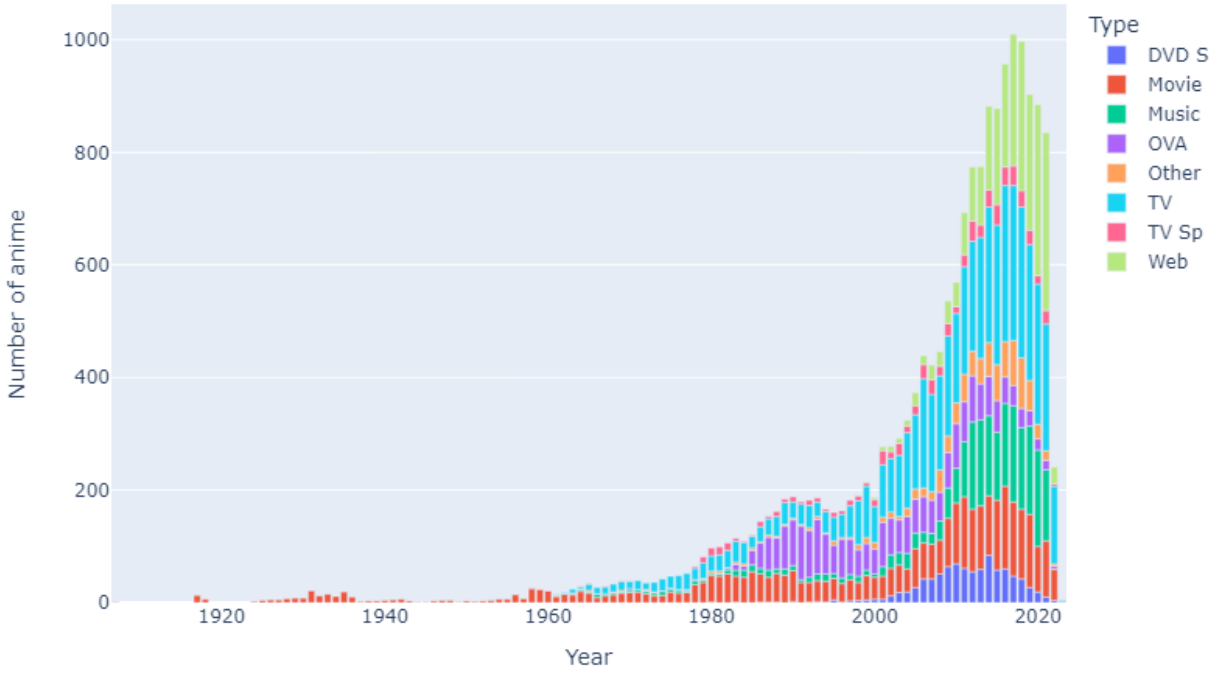

Figure 4: Ratings distribution

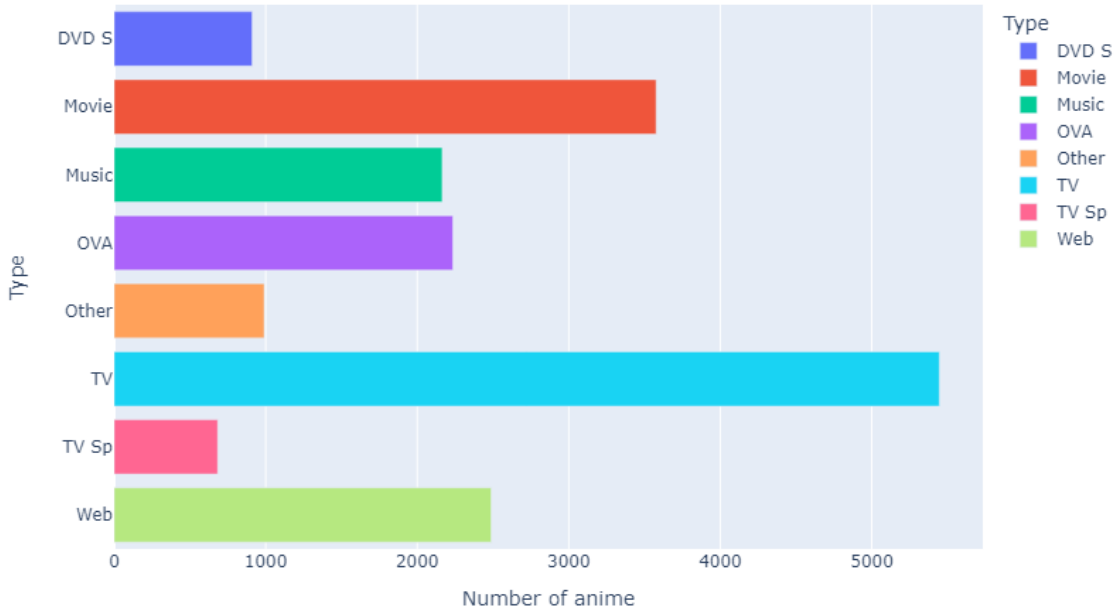

Figure 5: All of the anime types

So how about aggregating all data from the beginning of the anime industry to see which type is produced the most? By doing a simple `group` operation on the data frame and plotting the newly obtained data, Figure 5 is produced. As expected, studios prefer TV series anime the most when it comes to production. Movies are also very popular as ticket sales are also one of the main sources of profits for studios. Even though web-based and music anime are just getting their popularity in the past 5 years, they already have more than 2000 shows produced.

5 The most popular anime genre

The genre of the anime plays an important role when it comes to understanding audience preferences. In this section, I will use tags and genres interchangeably. `Tag` attribute in the data set is somewhat structured as it contains all of the genres in the anime separated by commas, hence, I just need to split the tags string in each entry to form arrays of tags enabling easier calculation. After summing up every tag, the top 10 most popular genre is shown in Figure 6.

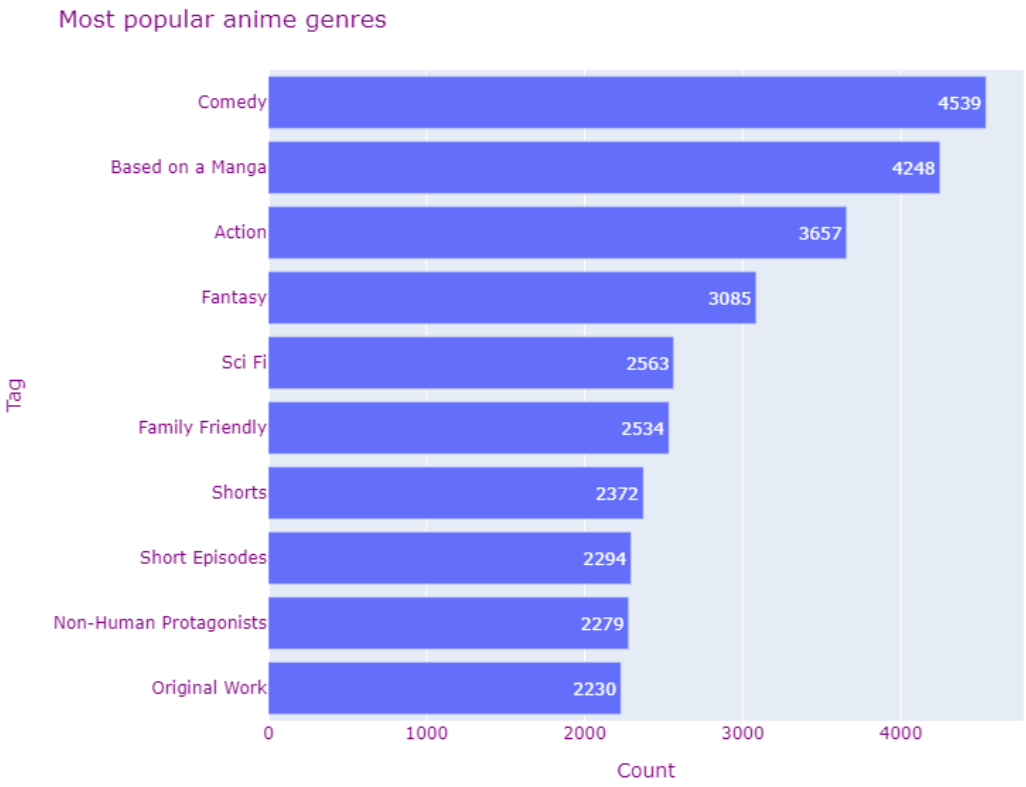

Figure 6: Top 10 most popular genres of all time

Immediately from the plot, comedy is the most popular tag in the anime industry, suggesting that the audience enjoys watching comedy as a way to relax and de-stress. The anime adaptation of manga is also one of the main sources of inspiration studios use to produce animation and most of the shows are directly based on existing manga. This has two major pros: 1) Existing manga already has a fairly big fan base, so it is economically efficient to have an animated version of it. 2) It could be used as an advertisement for a new comic series, which is a common practice in the industry. However, this chart aggregates data from the early days of the anime industry since 1907, so it might not represent the current audience interests. Figure 7 shows the popularity in the past 10 years.

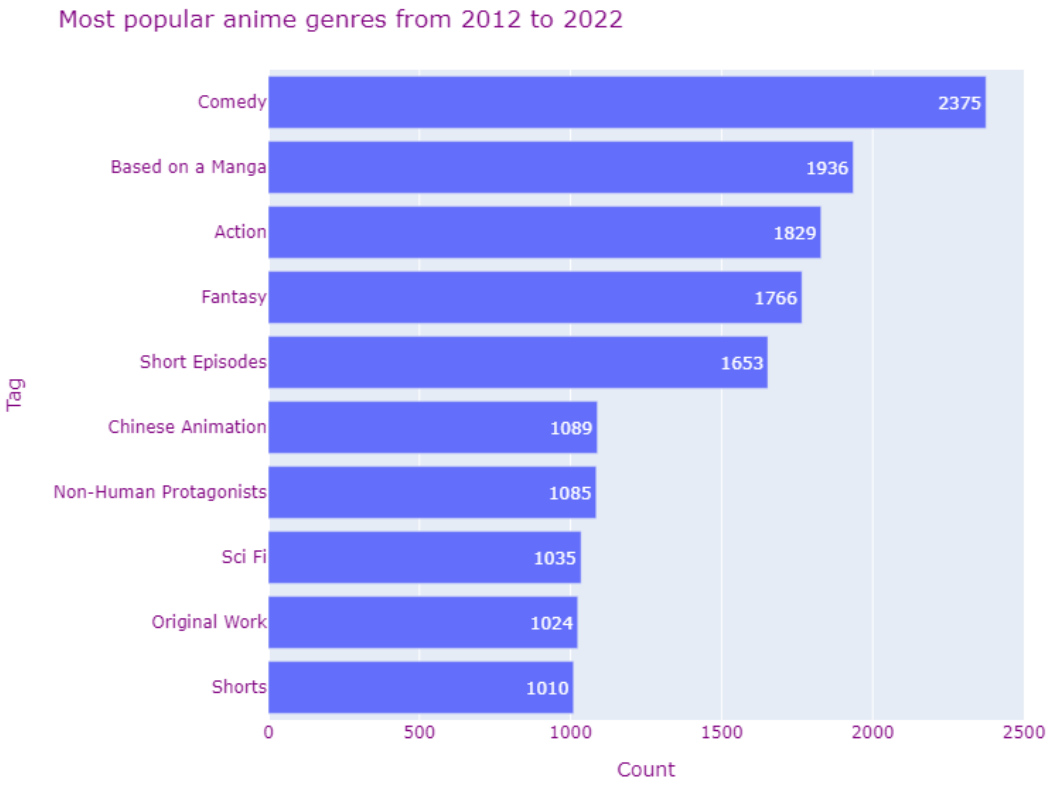

Figure 7: Top 10 most popular genres from 2012 to 2022

Still, the popularity of comedy still makes it the most produced genre in recent years. Original work in anime disappears from the top 10, suggesting that more and more manga or novel adaptation of anime is now the current trend when it comes to production. Surprisingly, Chinese Animation is now at the top 6 with more than a thousand shows have been produced. This shows that the animation industry in China has been growing rapidly in recent years.

6 The best anime studio

In this section, I will investigate which studio is the best producer of anime. The metric for this evaluation is quite simple, The average ratings from all of the produced anime. Considering that the number of shows affects the value, I only consider studios producing more than 50 anime. Since the Studio attribute of the data set is the most structured data, it is really easy to do computation on it. I just need to filter out all studios with less than 50 anime and calculate the average rating of each studio. The result is Figure 8. The figure shows that WIT Studio has the highest ratings from the audience. So what genres does this studio mainly produce? To visualize the genre, I use WordCloud [9] library in Python. We can easily see that the studio mainly covers anime adaptations of manga. The main theme of the studio is pretty dark such as Post-apocalyptic, Drama, Violence, and Horror. The studio also focuses on Action movies, and one of the most famous anime they produced is Attack on Titan, which claims two spots on the most-rated anime of all time in section 3.

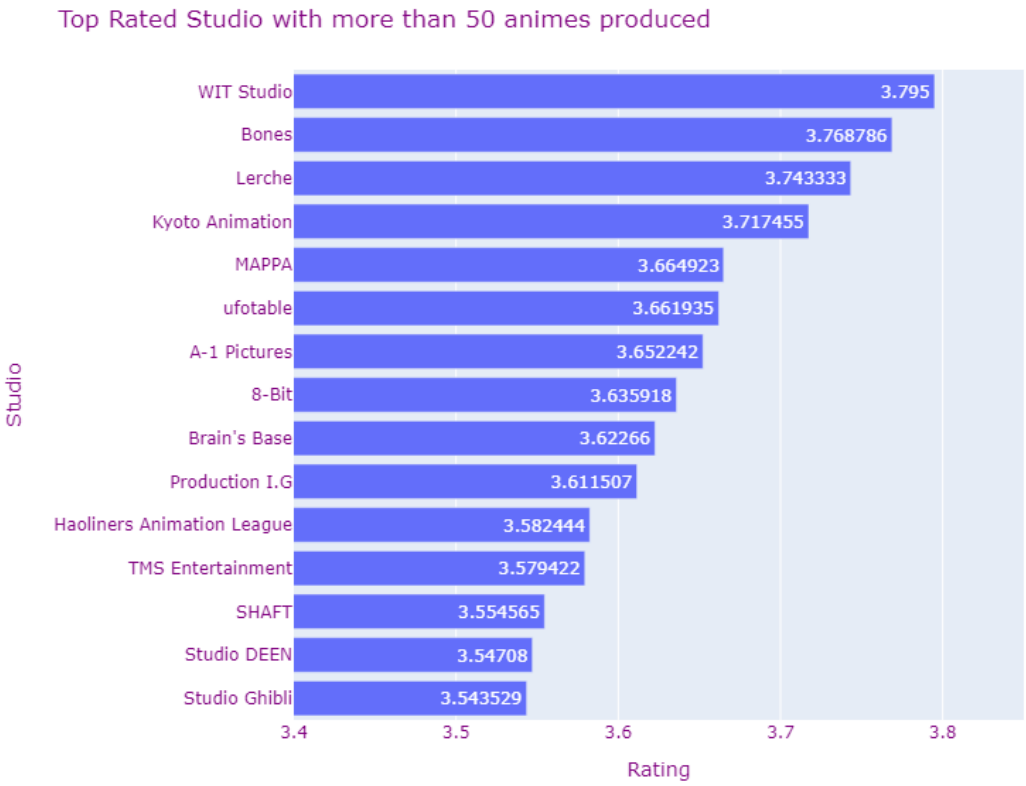

Figure 8: Top 10 most popular genres from 2012 to 2022

Figure 9: WIT Studio's genre

References

[1] https://www.kaggle.com/datasets/vishalmane10/anime-dataset-2022

[2] https://www.anime-planet.com/

[3] https://numpy.org/

[4] https://pandas.pydata.org/

[5] https://matplotlib.org/

[6] https://www.kaggle.com/datasets/vishalmane10/anime-dataset-2022

[7] https://requests.readthedocs.io/en/latest/

[8] https://beautiful-soup-4.readthedocs.io/en/latest/

[9] https://www.wordclouds.com/