

Project 1: An Exploratory Data Analysis on Anime

Daniel Hall

Rob Maysent

Ryan Johnson

SMU Bootcamp of 2023

Introduction

For Project 1, we chose to do our exploratory data analysis on Anime. We explored the correlations between subjects such as production studio ratings, the most popular genres, and the popularity of shows by year. We went in-depth over the analysis of anime to convey an intriguing story across these determined topics.

If you are unfamiliar, you may be asking yourself, “What exactly is Anime?” Anime is a style of animation that originated in Japan and has a distinct Japanese style. The term anime is derived from the English word animation, and in Japan, it is used to describe all forms of animation, regardless of origin and style. Anime is a combination of graphic art, characterization, cinematography, and other creative methods. In other words: all anime shows are cartoons, but not all cartoons are anime. While anime may have its roots in Japan, it has since spread to other countries and can be found in many different languages, including English. These styles of shows have gained widely received popularity in the Western region over the past two decades.

With the background understood, we wanted to take a deeper look at the data and get an understanding of the overall queries we had. “If someone is new to the concept of anime, what genres would be the best option to offer, and subsequently what production companies should they check out?” Or “If we wanted to write a new successful Anime show, which genres would give the best audience reception, and which production studios should we work with to give it the best shot possible?” We were left with 3 primary questions:

1. Which anime studio has the most top-rated anime?
2. What are the most popular genres?
3. Which anime was the most popular by the year, and are there any trends in the top shows?

This led us to our hypothesis: As an anime studio produces more shows, the quality or popularity of fan perception will gradually decline.

Data Cleaning

An issue we quickly came across was the fact that most data files we had access to were limited in scope and did not include all the information we were hoping to find. For example, we could not ascertain information regarding revenues gained from production companies, nor could we get any insight as to who the main viewing audience was, which would include items such as age, sex, region, etc. These are privately owned statistics that we just could not come across without some sort of clearance. With that being said, we were able to settle on two datasets that still provided an insightful amount of information. One we labeled 'main Data Frame' and the other 'studio Data Frame.' [Fig. 1.1 & 1.2]

We first proceeded with importing our libraries followed by reading the csv files. Before merging the two datasets we finalized, we needed to make sure our columns matched. To do this, we chose to change "title" to "name" in our 'main Data Frame.' [Fig. 1.3] Once complete, we could merge the data sets based on the names of the anime shows. There were three columns labeled as "Unnamed" in the 'main Data Frame' that were not needed, so we chose to drop them to present a more accurate and concise data set. [Fig. 1.4 & 1.5] Then we had to split strings in our "aired" column to create two new columns to represent a "Start Date" & "End Date." [Fig. 1.6] We performed a split for the genre as well. [Fig. 1.7] Our genre column had concatenated strings, so we had to run a code in which each row in the genre would only have the first classification listed. We believe this was best to convey the true (or dominant) category of a show. In addition, our data sets included both TV shows and Movies, so we needed to create a mask to keep only TV shows. This was determined as the best way to convey an undiluted bias when it comes to ratings and popularity. You may have one studio that makes only one movie (on its own or with collaboration) and be a one-hit-wonder or a major flop. This would lead to an outlier and/or skew in the data. We ended up with the following columns: "uid, Name, genre, episodes, members, popularity, ranked, score, Type, Studio, Start Date, End Date."

Anime Studios with the Most Top-Rated Anime

Starting with our first question, we needed to find out which anime studio produced the most top-rated anime TV shows. This was best conveyed as a bar chart. [Fig. 2.1] To narrow down our scope we chose to keep the shows with ratings of an 8 or higher. It is shown that *Madhouse* has the most success when it comes to ratings, boasting a total of 70 shows with a ranking of 8 or higher. They are a Japanese animation studio founded in 1972 that has produced numerous anime popular amongst fans, such as *HunterXHunter*, *Death Note*, and *One Punch Man*.

To show the extent of how one-of-a-kind this situation is, we provided a scatter plot displaying the “Average Rating by Amount of Anime Show.” [Fig. 2.2] This plot shows the average rating amongst production studios by the number of shows released. With an emphasis on *Madhouse*, we can see most other studios produced at most 20 shows and had an average rating of 7 to 8 in comparison. While some studios had higher ratings, they did not produce as many shows. This begs the question of whether they would have seen continued success the more they put out, or if they would see a downfall with an oversaturated market distribution. *Madhouse* was able to both put out numerous shows, an immense amount of 108 shows as well as maintain a higher member rating.

To further investigate the success of *Madhouse*, we began to consider what genres they produced the most. With a grouped bar chart, we can compare 5 studios based on genre and the number of shows they produced in each genre. [Fig. 2.3] The genres chosen were “Action, Adventure, Comedy, Mystery, and Slice of Life.” *Madhouse* appeared to not stick to one specific genre, while others did. They are completely missing from ‘Mystery’ and do not release many ‘Comedy’ shows, whereas *J.C. Staff*, the studio with the second highest number of top-rated shows, primarily focused on this genre. This could be why *Madhouse* has success due to how versatile they are.

Most Popular Genres in Anime

Just as colorful as the shows themselves, the same is the spectrum of genres associated with anime. To find if one genre is preferred over the others, we produced a bar chart for comparison. [Fig. 3.1] This will show the “Top 15 Genres by Frequency”. This list included “Action, Comedy, Hentai, Adventure, Music, Slice of Life, Kids, Drama, Fantasy, Sci-Fi, Dementia, Historical, Mystery, Sports, Other.” The ‘Action’ genre took first place with 4,120 anime appearances in our data set after cleaning. The ‘Other’ category was a conglomerate of all other genres available, with most having less than 200 appearances. A table was presented to show all genres with a quantity above 10 appearances. [Fig. 3.2]

To further break down genre categories by quantity, we wanted to compare the top two genres to the others on the list. A pie chart was set up to indicate our top genres as a part of a whole. [Fig. 3.3] There is a steep difference between those at the top of the list and those at the tail end. These remaining few still hold their place in the list as many were listed as 'sub-categories' in the original data sets before the data cleaning. One genre accounted for 25.1% of the top categories, showing that a fourth of all shows produced can be classified under ‘Action.’ The ‘Comedy’ genre came in second with 20.3% (or a fifth) of all classifications. Given the pure nature of anime, we originally assumed genres, such as Fantasy and Superpower, would sit higher on the list. Some genres we did not necessarily expect to be most common were Music (at 7.9%) and Slice of Life (at 5.9%). This assumes there is a larger fan base for slower, more casual viewers than expected. Even as avid anime watchers ourselves, these are not ones we tend to lean towards, proving there is a show with classification and plot for everyone.

If most of the anime that gets produced is categorized under ‘Action,’ does this mean they are also the best-rated? Our next goal was to find a correlation between the most occurring genres and member ratings. Our ratings were on a scale of 1 to 10. The average genre rating came out to be 6.44 out of 10. ‘Action’ ranked 5th amongst the “The Top 10 Genres by Score” with a 6.84, while ‘Thriller’ takes 1st with an average score of 7.53. [Fig. 3.4] The ‘Josei’ (woman or female in Japanese) genre was close behind with a score of 7.52. With this latest information in mind, we wanted to compare our 2 top-rated genres to the 2 most common

genres to see if any outliers were affecting the outcome or if they share a correlation in the way of being true fan-favorite genres. Both 'Comedy' and 'Action' experienced some outliers affecting the median score, while 'Thriller' and 'Josei' remained consistent with their score votes. [Fig. 3.5]

Growth of Anime Through the Years

According to our data the first animation classified as 'Anime' was released in 1917, the growth of the industry since then has been explosive [Fig. 4.2]. The most popular genre over time is clearly indicated to be action, however a surprising increase in comedy is displayed as well. In fact, several years' data shows that action anime was the second most produced genre falling behind comedy multiple times before inevitably bouncing back [Fig. 4.1]. However, in 2020 the explosive growth of the medium was brought to a grinding halt as the coronavirus pandemic swept across Japan and the world. The effect on the anime industry was quickly apparent, with a staggering drop regarding the number of total shows produced in 2020 due to most production studios having no infrastructure for animators to work on shows from home. In the years since then the industry has recovered but our data does not cover anything after 2020 in its entirety.

Since we have such a historic dataset dating back to 1917, I wanted to illustrate what the longest running series are, so I found the longest running series by total episodes produced and the longest by time on the air. It turns out most of the longest running series are kids shows, with a few outliers of action series [Fig. 4.3 & 4.4]. I believe this to be caused by the most historic shows being originally produced for kids, it was not until close to the turn of the century that anime started being produced for adults on a normative basis. However, if we were to decide what the best genre to produce a show in based off this data it would come to be an action show for kids.

Regression

Initially looking at the data with the notion of needing to do a linear regression in mind, I first thought that there might be a considerable correlation between the number of total reviews left and that score then resultantly being higher. While that is the strongest correlation present with a correlation rating of .488, it is not great [Fig. 5.1 & 5.2]. However, for our purposes it will be a fine representation. The resultant R-squared values are extremely low for both the multiple linear regression and the simple linear regression, both of which provided near identical values of 0.191 and 0.19 respectively [Fig. 5.3 & 5.5].

Statistical Hypothesis

Given that anime is released on a seasonal basis regardless of genre, we hypothesized that all genres would share a similar airing span. To check this hypothesis, a 'difference' value was created to represent the days between the first airing date of a series and the respective finale, where possible, and compared the averages across genres. However, since we can see that the p-value of our t-tests is 0.02 and 0.03 [Fig. 6.1], respectively, for our comparisons of multiple genres, we must default to the alternate hypothesis, that there is a significant difference between the average airing time of genres.

Limitation & Bias

- The datasets we used had concatenated strings within our genres, so we had to do a split and only keep whatever genre was first under each row. This led to some genres not having enough data as well as some having too much data due to only relying on whatever genre was labeled first within the concatenated rows.
- In addition to the datasets used to chart the genre and rankings, we wanted to introduce another ranking system from a Western audience to contextualize popularity strictly here at home. Unfortunately, the naming scheme was different between rating

listings on the respective datasets, so we were forced to cut the western-specific rankings.

- For our data utilizing the start and end dates of a series, there is no denotation for ongoing series. This forced us to drop a sizable portion of data where there was no end date listed. This means there is missing data for some of the biggest shows, such as *One Piece* on graphs that require an end date.
- Like the previous limitation, a series that took an extended hiatus where there was no new material for long periods may have skewed the data, as they do not have a unique start and end date for each period where a series was active with new releases. Thus, there may be multiple series that have been marked as ongoing for multiple years with very few releases.

Future Work

Given the limitations we faced when gathering our data, we had to forego any comparisons of the viewing audiences. Had we been able to secure information on demographics (male or female, age range, etc.) we could have gained more insight into who exactly is the prime demographic for production studios.

Additionally, obtaining information regarding viewers per country and/or city would have provided us with an opportunity for Geo-mapping.

Additional questions to be answered:

- Which areas in the world have the highest density of viewers?
- Do certain locations prefer one genre over the other?
- Or has there been an uptick of viewers in the Western Regions within the last few decades as trends migrate?

Figures List Reference Page (Data Visuals by Appearance):

Figure 1.1

```
In [5]: main.head()
```

```
Out[5]:
```

	Unnamed: 0	uid	title	genre	aired	episodes	members	popularity	ranked	score
0	0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shoun...]	Oct 4, 2015 to Mar 27, 2016	25.0	489888	141	25.0	8.82
1	1	23273	Shigatsu wa Kimi no Uso	['Drama', 'Music', 'Romance', 'School', 'Shoun...]	Oct 10, 2014 to Mar 20, 2015	22.0	995473	28	24.0	8.83
2	2	34599	Made in Abyss	['Sci-Fi', 'Adventure', 'Mystery', 'Drama', 'F...]	Jul 7, 2017 to Sep 29, 2017	13.0	581663	98	23.0	8.83
3	3	5114	Fullmetal Alchemist: Brotherhood	['Action', 'Military', 'Adventure', 'Comedy', '...]	Apr 5, 2009 to Jul 4, 2010	64.0	1615084	4	1.0	9.23
4	4	31758	Kizumonogatari III: Reiketsu-hen	['Action', 'Mystery', 'Supernatural', 'Vampire']	Jan 6, 2017	1.0	214621	502	22.0	8.83

Figure 1.2

```
In [6]: studio.head()
```

```
Out[6]:
```

	Unnamed: 0.1	Unnamed: 0	Name	Type	Studio	Genres
0	0	0	Fullmetal Alchemist: Brotherhood	TV	['Bones']	['Action', 'Adventure', 'Drama', 'Fantasy']
1	1	1	Spy x Family	TV	['Wit Studio', 'CloverWorks']	['Action', 'Comedy']
2	2	2	Shingeki no Kyojin Season 3 Part 2	TV	['Wit Studio']	['Action', 'Drama']
3	3	3	Steins;Gate	TV	['White Fox']	['Drama', 'Sci-Fi', 'Suspense']
4	4	4	Gintama*	TV	['Bandai Namco Pictures']	['Action', 'Comedy', 'Sci-Fi']

Figure 1.3

```
In [14]: main = main.rename(columns = {"title": "Name"})
```

```
In [16]: df = pd.merge(main, studio, how = "left", on = ["Name", "Name"])
```

Figure 1.4

	Unnamed: 0_x	uid	Name	genre	aired	episodes	members	popularity	ranked	score	Unnamed: 0.1	Unnamed: 0_y	Type	Studio	Genres
0	0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shoun...']	Oct 4, 2015 to Mar 27, 2016	25.0	489888	141	25.0	8.82	63.0	63.0	TV	[Production I.G]	['Sports']
1	1	23273	Shigatsu wa Kimi no Uso	['Drama', 'Music', 'Romance', 'School', 'Shoun...']	Oct 10, 2014 to Mar 20, 2015	22.0	995473	28	24.0	8.83	57.0	57.0	TV	[A-1 Pictures]	['Drama', 'Romance']
2	2	34599	Made in Abyss	['Sci-Fi', 'Adventure', 'Mystery', 'Drama', 'F...']	Jul 7, 2017 to Sep 29, 2017	13.0	581663	98	23.0	8.83	52.0	52.0	TV	[Kinema Citrus]	['Adventure', 'Drama', 'Fantasy', 'Mystery', '...']
3	3	5114	Fullmetal Alchemist: Brotherhood	['Action', 'Military', 'Adventure', 'Comedy', '...']	Apr 5, 2009 to Jul 4, 2010	64.0	1615084	4	1.0	9.23	0.0	0.0	TV	[Bones]	['Action', 'Adventure', 'Drama', 'Fantasy']
4	4	31758	Kizumonogatari III: Reiketsu-hen	['Action', 'Mystery', 'Supernatural', 'Vampire']	Jan 6, 2017	1.0	214621	502	22.0	8.83	29.0	29.0	Movie	[Shaft]	['Action', 'Mystery', 'Supernatural']

Figure 1.5

```

In [78]: df.columns
Out[78]: Index(['Unnamed: 0_x', 'uid', 'Name', 'genre', 'aired', 'episodes', 'members',
               'popularity', 'ranked', 'score', 'Unnamed: 0.1', 'Unnamed: 0_y', 'Type',
               'Studio', 'Genres'],
              dtype='object')

In [79]: cols = ['uid', 'Name', 'genre', 'aired', 'episodes', 'members',
               'popularity', 'ranked', 'score', 'Type',
               'Studio', 'Genres']
df = df.loc[:, cols]

In [80]: df.head()
Out[80]:

```

	uid	Name	genre	aired	episodes	members	popularity	ranked	score	Type	Studio	Genres
0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shoun...']	Oct 4, 2015 to Mar 27, 2016	25.0	489888	141	25.0	8.82	TV	[Production I.G]	['Sports']
1	23273	Shigatsu wa Kimi no Uso	['Drama', 'Music', 'Romance', 'School', 'Shoun...']	Oct 10, 2014 to Mar 20, 2015	22.0	995473	28	24.0	8.83	TV	[A-1 Pictures]	['Drama', 'Romance']
2	34599	Made in Abyss	['Sci-Fi', 'Adventure', 'Mystery', 'Drama', 'F...']	Jul 7, 2017 to Sep 29, 2017	13.0	581663	98	23.0	8.83	TV	[Kinema Citrus]	['Adventure', 'Drama', 'Fantasy', 'Mystery', '...']
3	5114	Fullmetal Alchemist: Brotherhood	['Action', 'Military', 'Adventure', 'Comedy', '...']	Apr 5, 2009 to Jul 4, 2010	64.0	1615084	4	1.0	9.23	TV	[Bones]	['Action', 'Adventure', 'Drama', 'Fantasy']
4	31758	Kizumonogatari III: Reiketsu-hen	['Action', 'Mystery', 'Supernatural', 'Vampire']	Jan 6, 2017	1.0	214621	502	22.0	8.83	Movie	[Shaft]	['Action', 'Mystery', 'Supernatural']

Figure 1.6

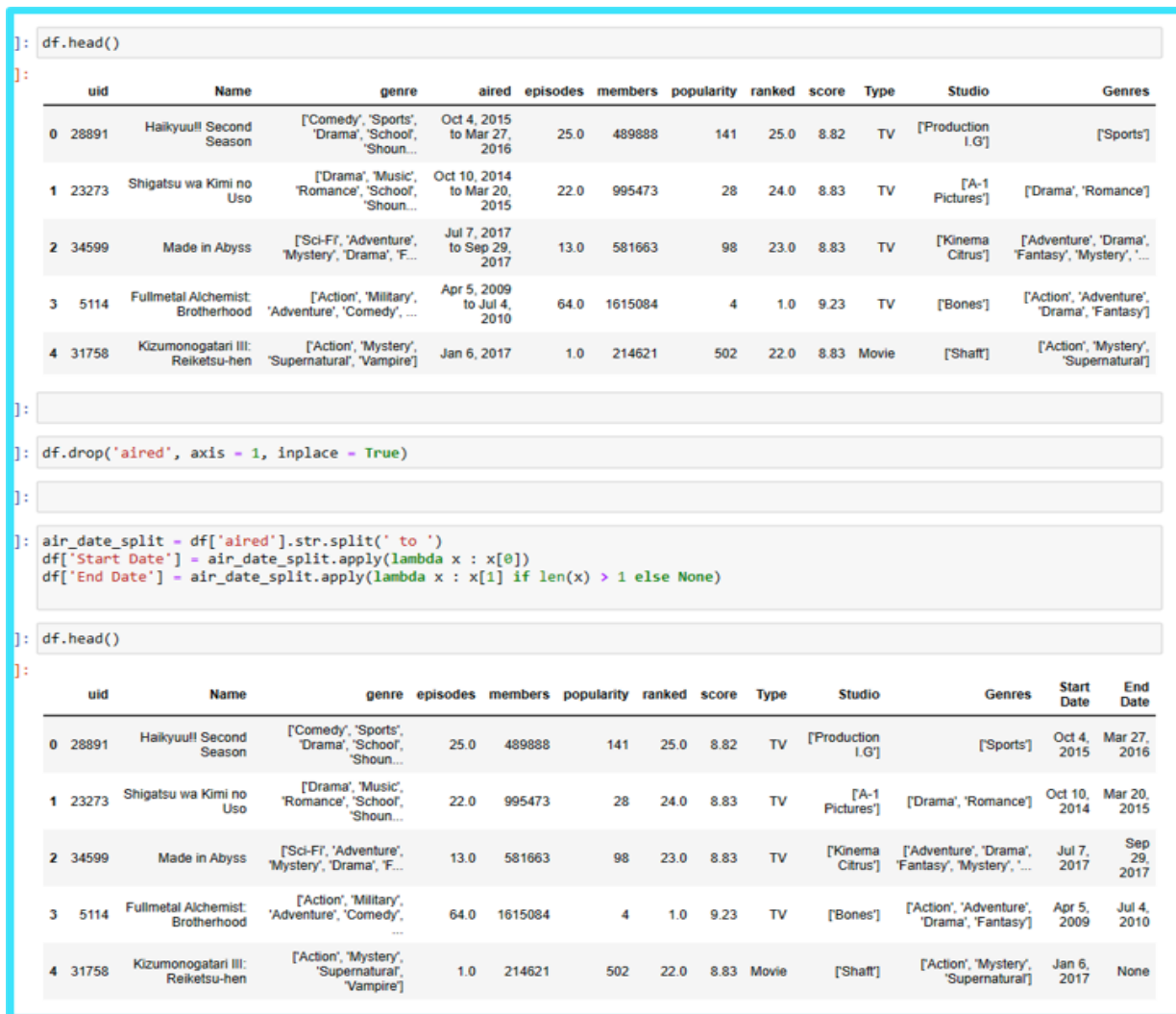


Figure 1.7

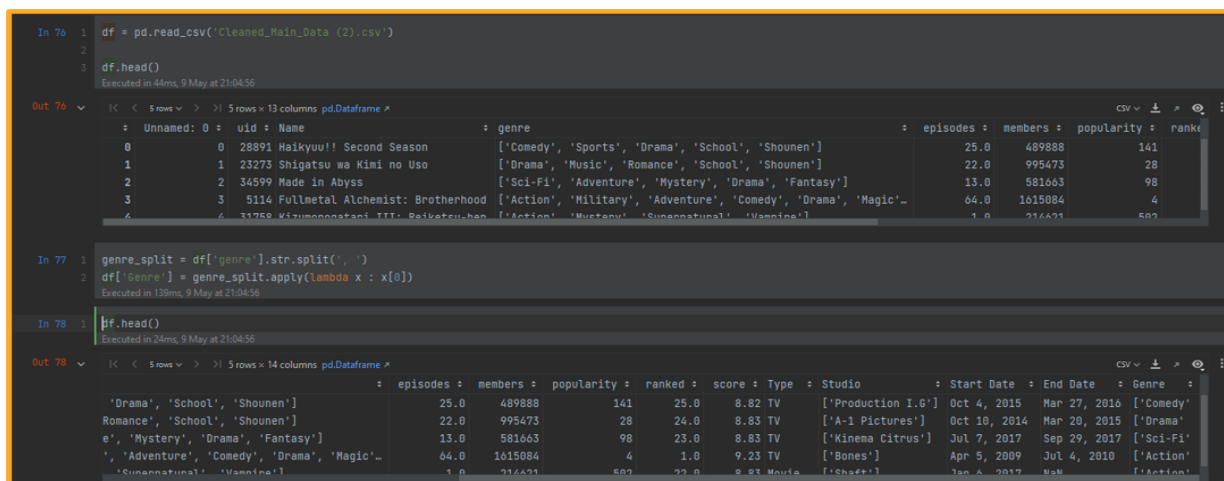


Figure 2.1

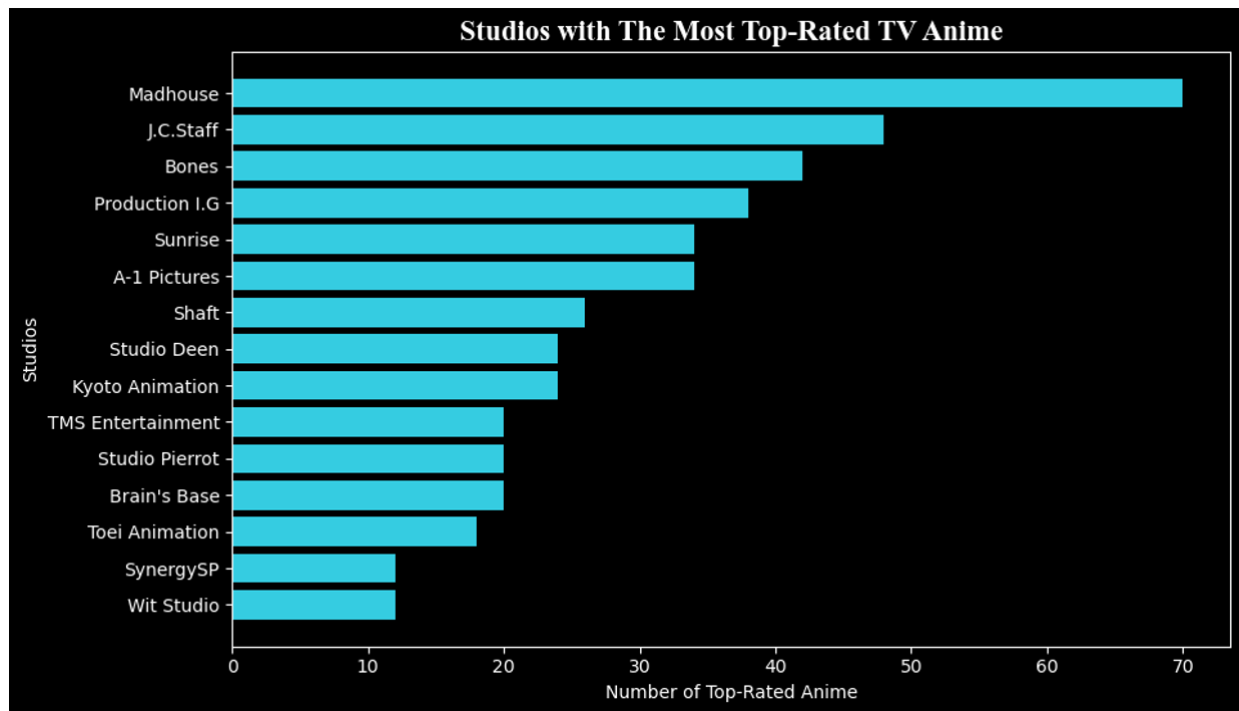


Figure 2.2

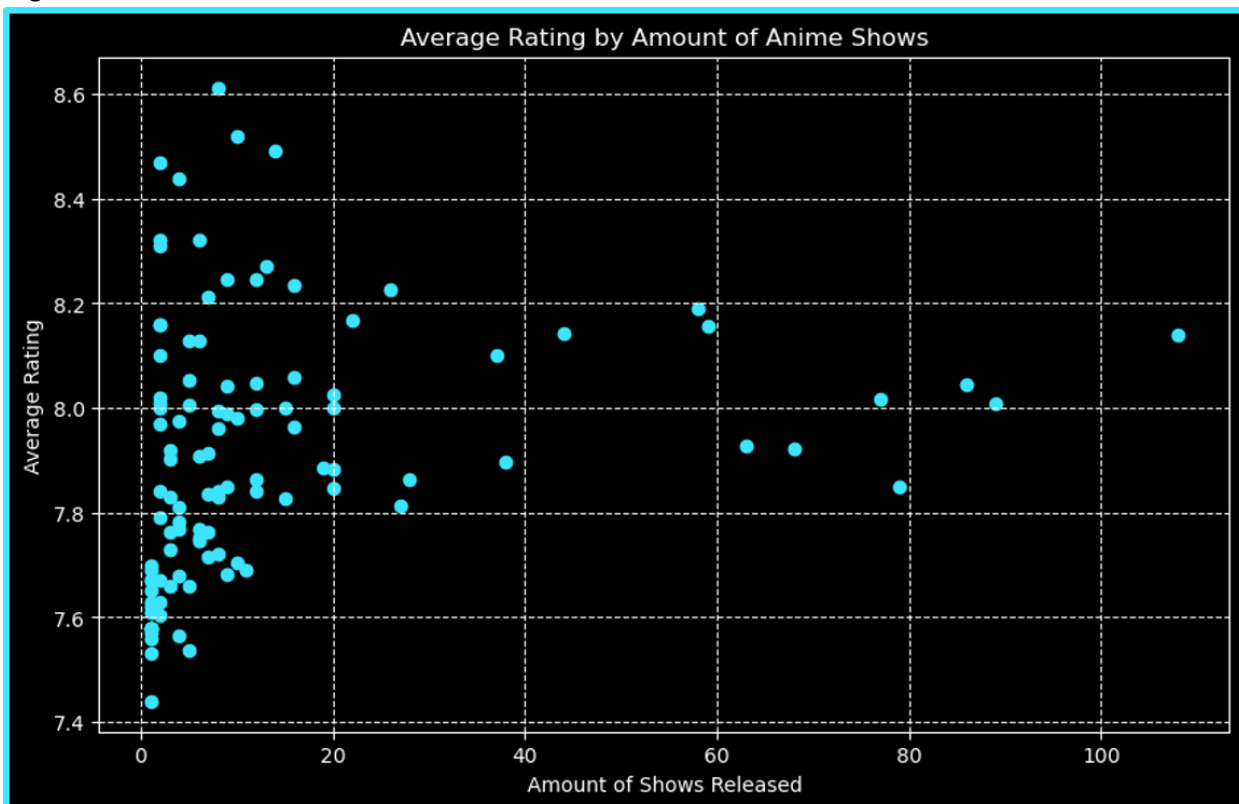


Figure 2.3

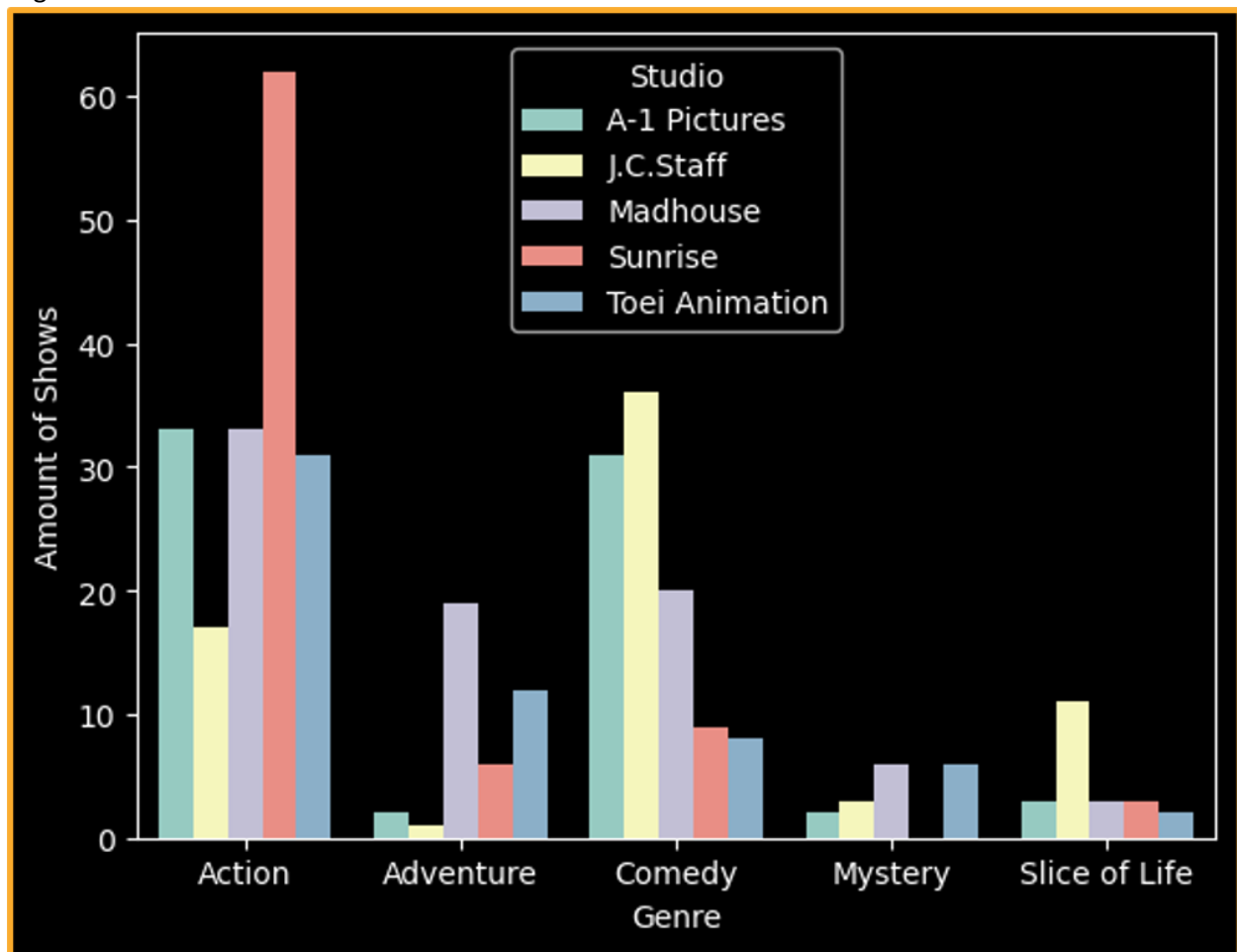


Figure 3.1

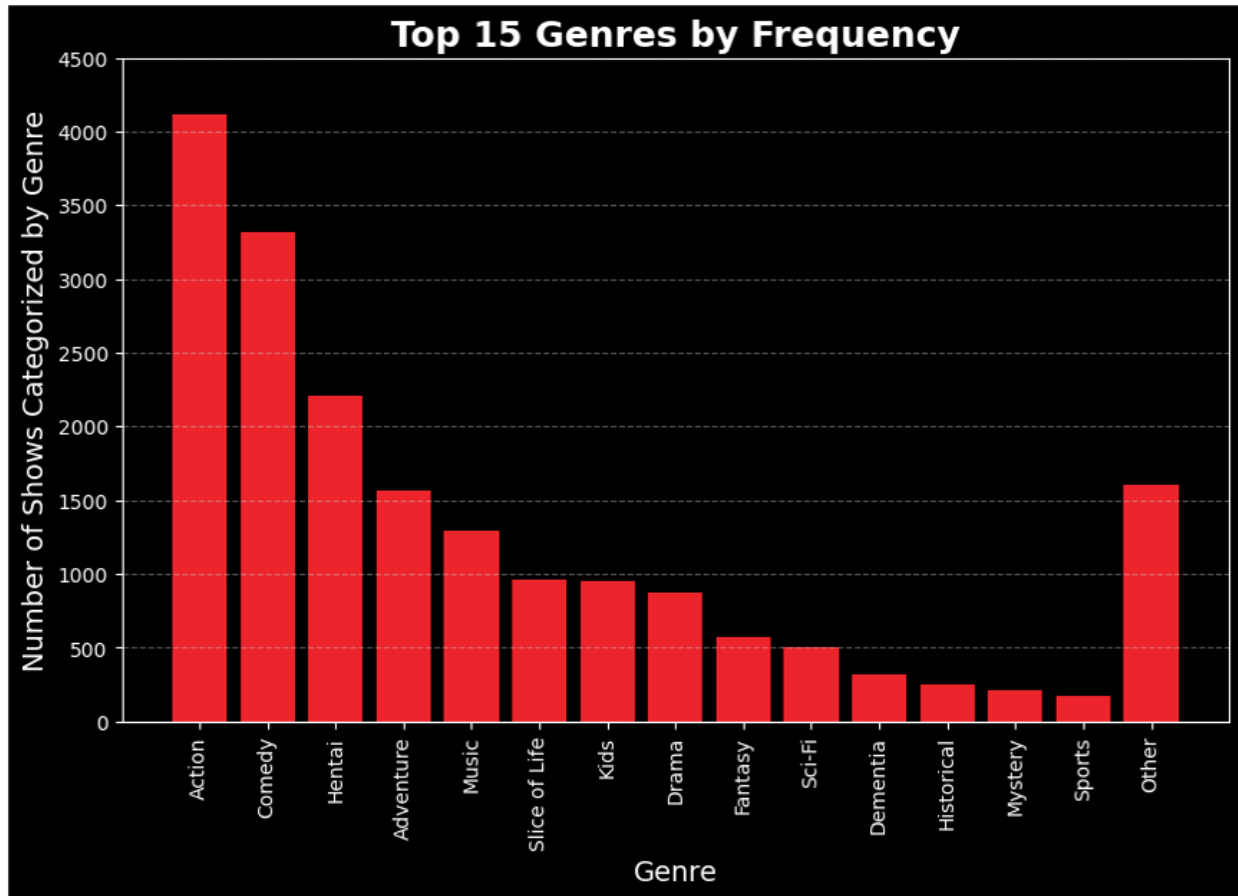


Figure 3.2

Action	4120
Comedy	3325
Hentai	2212
Adventure	1567
Music	1296
Slice of Life	963
Kids	948
Drama	876
Fantasy	573
Sci-Fi	504
Dementia	317
Historical	247
Mystery	209
Sports	177

Game	176
Romance	158
Harem	156
Ecchi	146
Magic	112
Military	108
Mecha	82
Demons	82
Horror	78
Parody	78
Cars	60
Supernatural	56
School	52
Psychological	49
Space	36
Shounen	21
Police	14
Super Power	11
Seinen	11

Figure 3.3

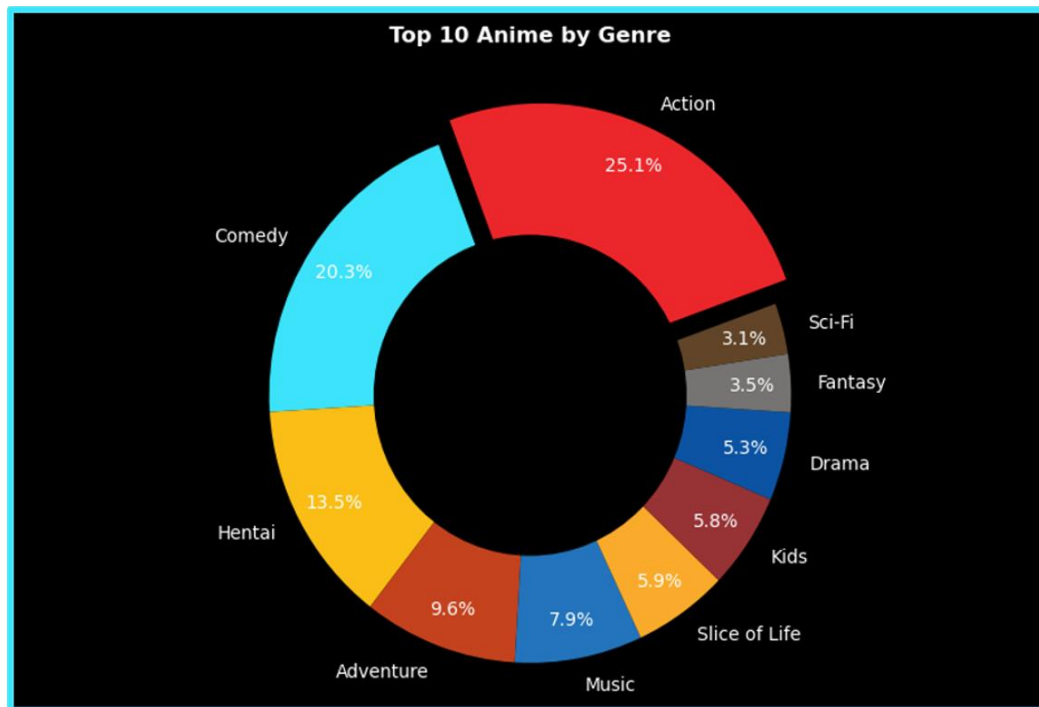


Figure 3.4

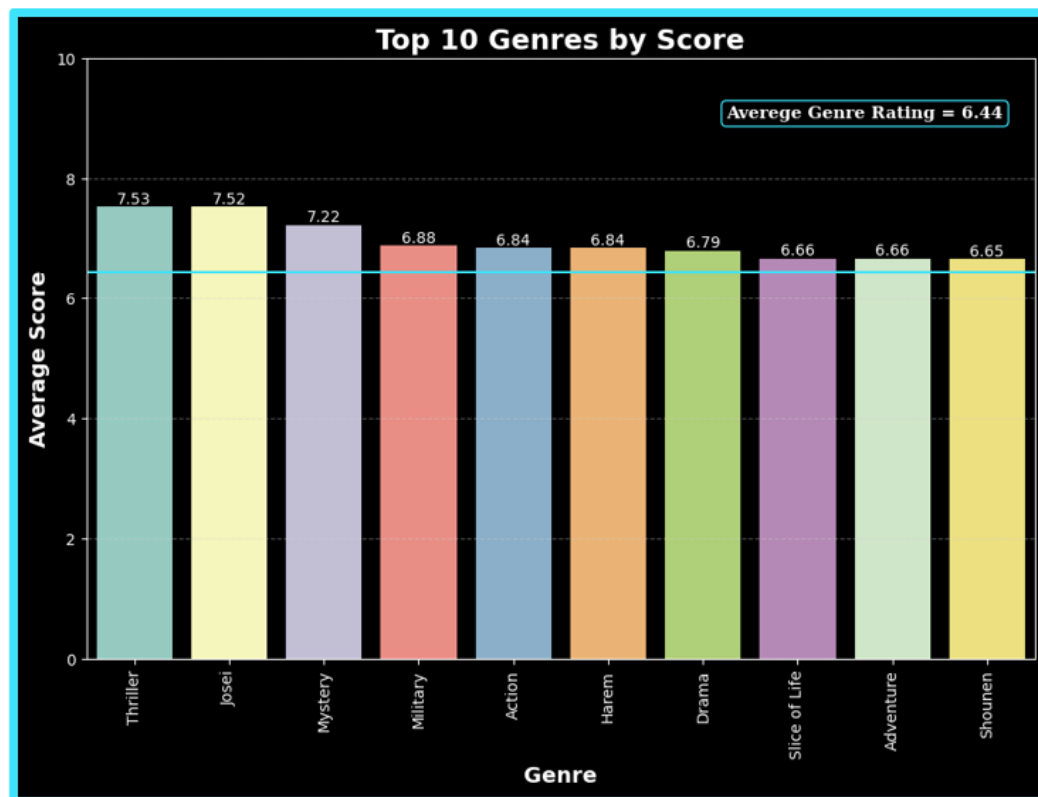


Figure 3.5

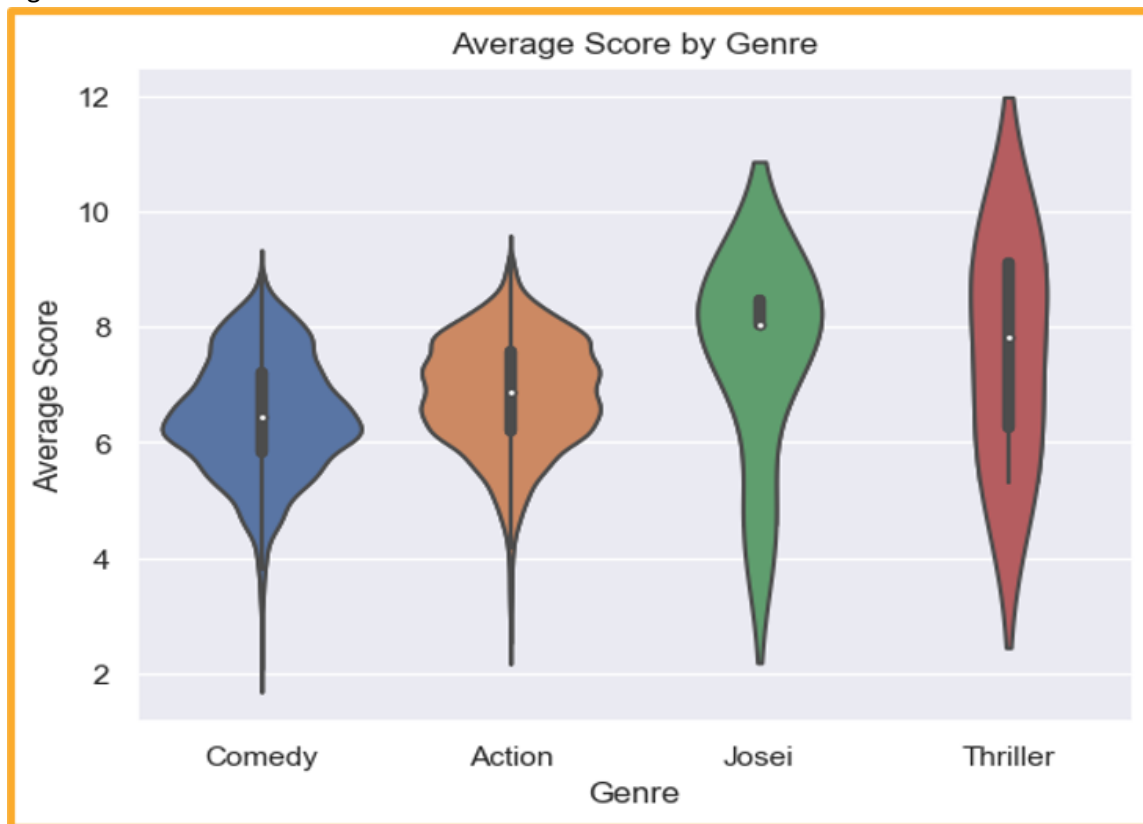


Figure 4.1

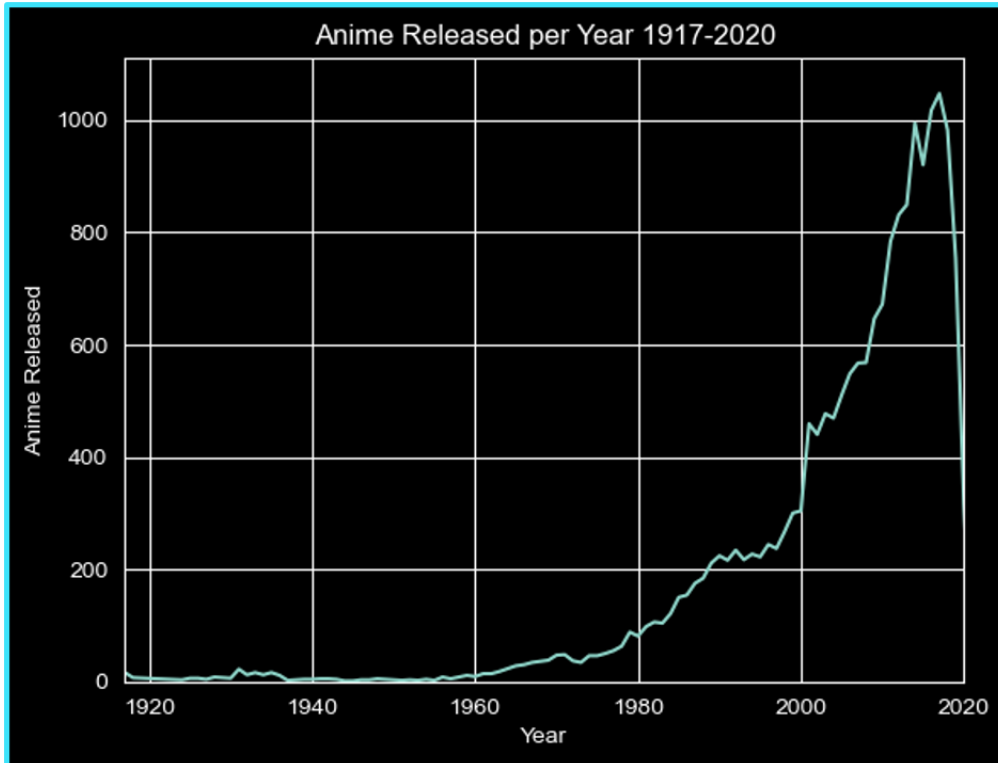


Figure 4.2

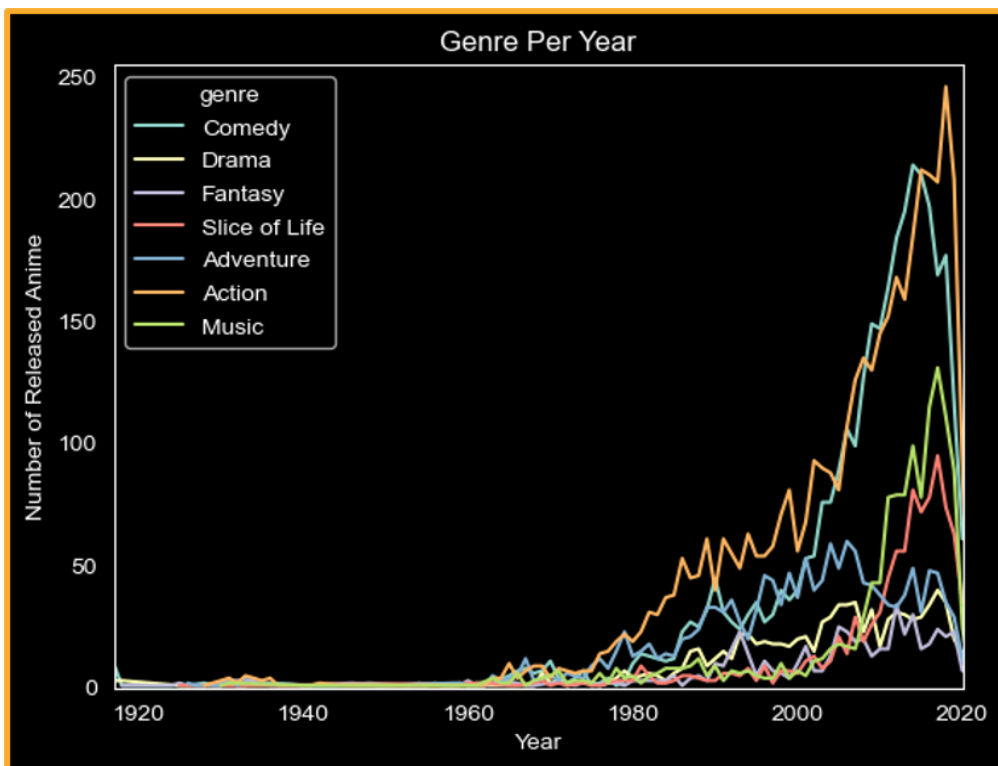


Figure 4.3

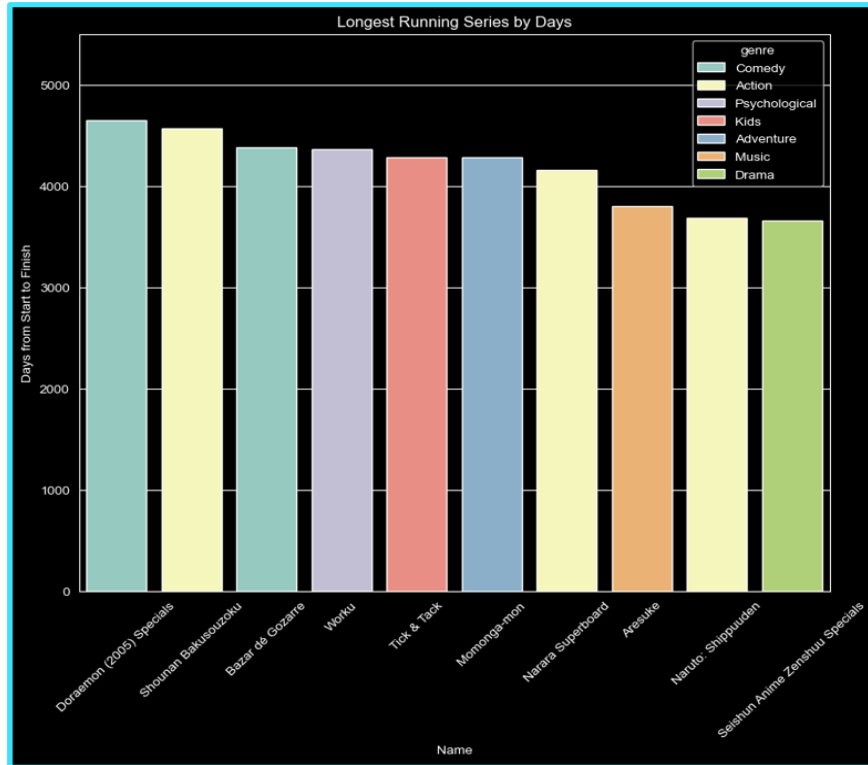


Figure 4.4

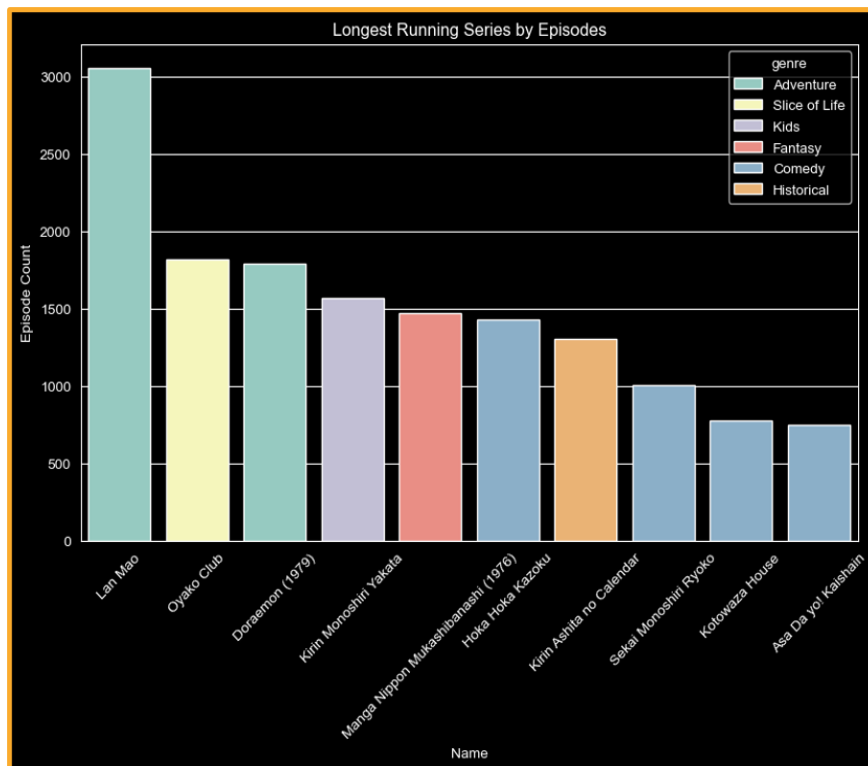
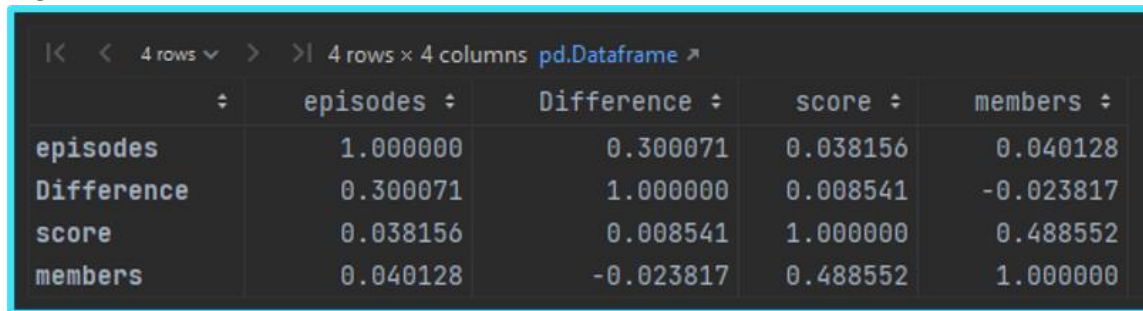


Figure 5.1



	episodes	Difference	score	members
episodes	1.000000	0.300071	0.038156	0.040128
Difference	0.300071	1.000000	0.008541	-0.023817
score	0.038156	0.008541	1.000000	0.488552
members	0.040128	-0.023817	0.488552	1.000000

Figure 5.2

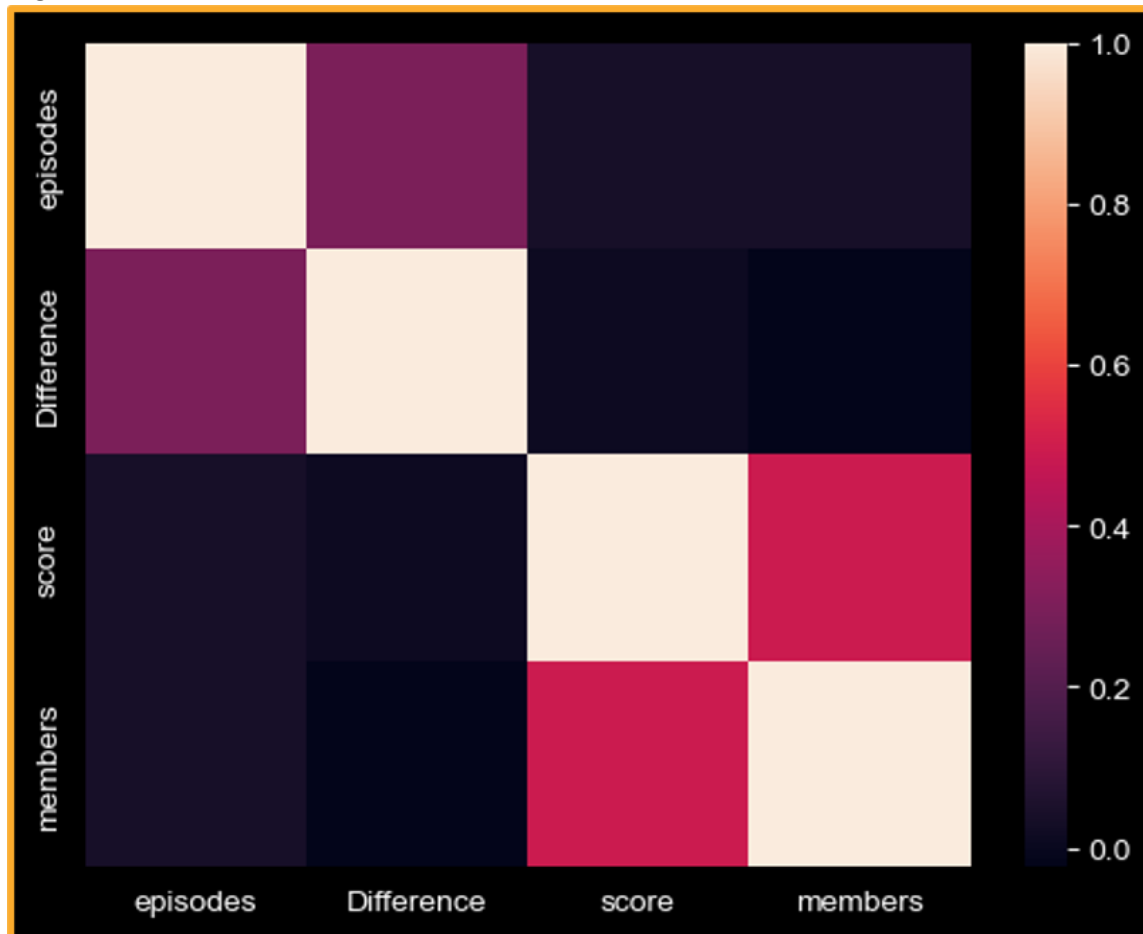


Figure 5.3

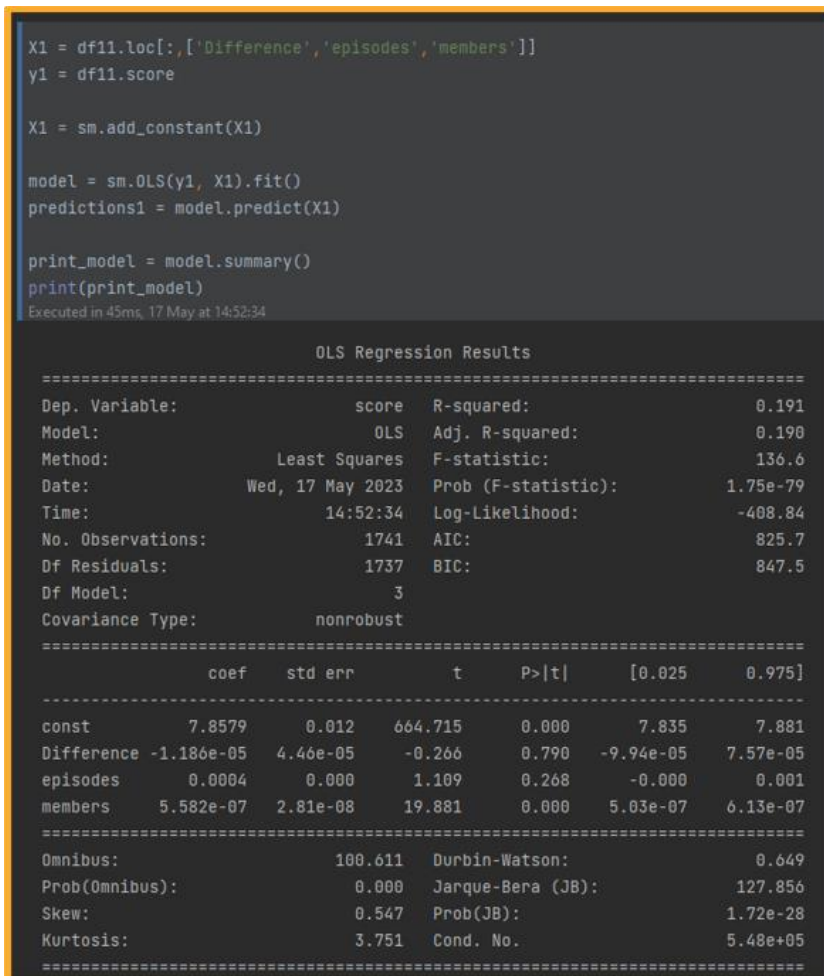


Figure 5.4

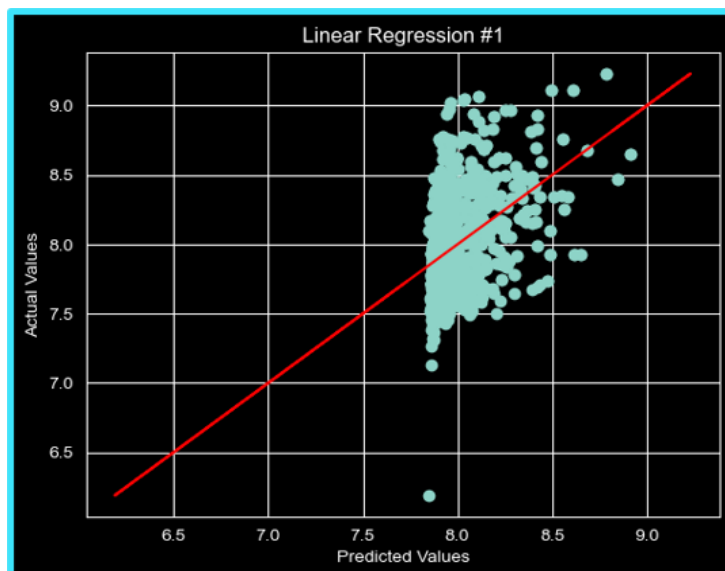


Figure 5.5

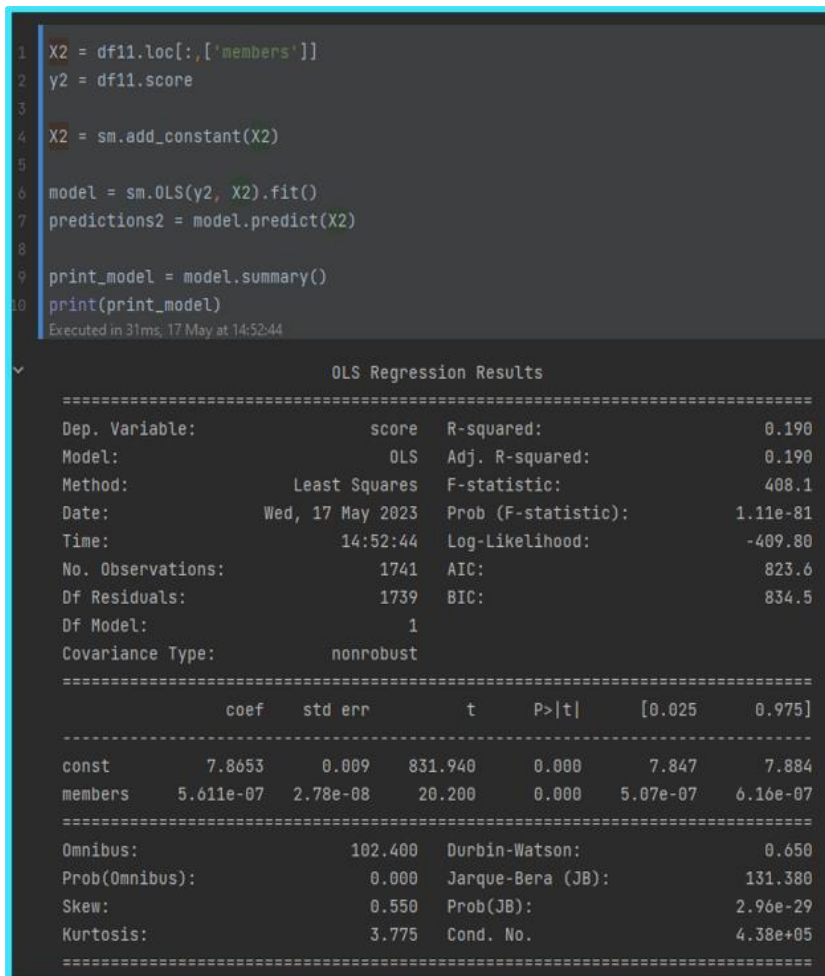


Figure 5.6

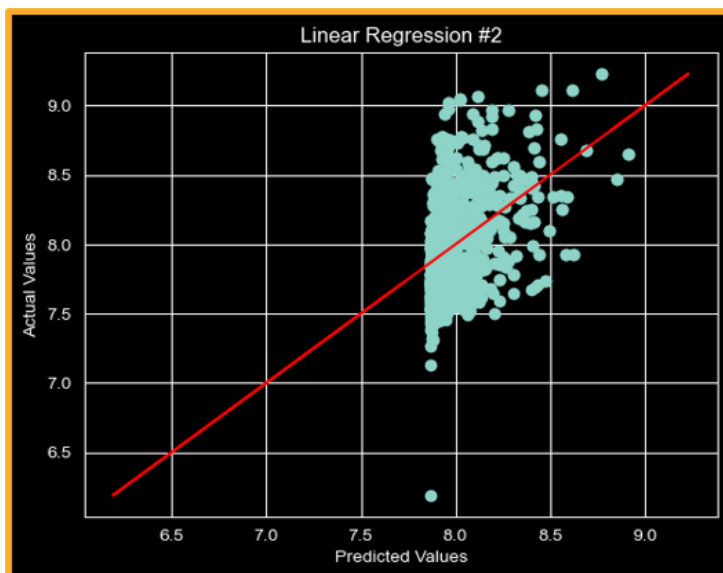


Figure 6.1

```
1 group1 = df11[df11['genre']=='Music']
2 group2 = df11[df11['genre']=='Comedy']
3
4 ttest_ind(group1['Difference'],group2['Difference'])
   Executed in 24ms, 17 May at 14:48:41

   Ttest_indResult(statistic=2.197314848205368, pvalue=0.028511491895022962)

1 group1 = df11[df11['genre']=='Drama']
2 group2 = df11[df11['genre']=='Action']
3
4 ttest_ind(group1['Difference'],group2['Difference'])
   Executed in 7ms, 17 May at 14:46:26

   Ttest_indResult(statistic=-2.142553261574121, pvalue=0.03247112738624942)
```

References:

The first dataset has over 2,000 anime listed while the second dataset was supplemental, as we needed a column for which studios made individual anime.

- <https://www.kaggle.com/datasets/angadchau/anime-dataset>
- <https://www.kaggle.com/datasets/brunobacelardc/myanimelist-top-1000-anime>