

EXPLORATORY DATA ANALYSIS ON ANIME

DANIEL HALL
ROB MAYSENT
RYAN JOHNSON



Table of Contents:

Introduction to Exploratory Data Analysis On Anime

Research Questions

Our Datasets Used

Data Cleaning

The Most Popular Studios

The Most Popular Genres

The Growth of the Industry

Linear Regression & Statistical Hypothesis

Conclusion

TABLE OF CONTENTS



INTRODUCTION

Today we will be going in-depth over the analysis of anime to convey an intriguing story across determined topics, such as the numerous animation studios that make anime, the different genres within anime, as well as attempt to find a pattern when it comes to viewers that may watch a specific genre.

You may have heard of anime but what exactly is it? Anime is a hand-drawn or computer-generated animation that originated in Japan but has gained popularity in the Western region over the past two decades.

In other words: all anime shows are cartoons, but not all cartoons are anime.



RESEARCH QUESTIONS

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

Hypothesis: As an anime studio produces more shows, the quality or popularity from fan perception will gradually decline.



OUR DATASETS

- The two datasets that we used were sourced from Kaggle:
<https://www.kaggle.com/datasets/angadchau/anime-dataset>
<https://www.kaggle.com/datasets/brunobacelardc/myanimelist-top-1000-anime>
- The first dataset has over 2000 anime listed, whereas the second dataset was supplemental, as we needed a column for which studios made individual anime.
- Our priority was to clean up the data and merge them by anime name to have a complete Data Frame so we could evaluate our data properly and answer our research questions.

DATA CLEANING

RENAMING & MERGING DATA FRAMES

```
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

Step 1

```
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']

Step 2

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

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

Step 3

Before merging the two data sets, we needed to make sure our columns matched in our 'main Data Frame' and 'studio Data Frame'. To do this, we chose to change "title" to "name" in our 'main Data Frame'. Once complete, we could merge the data sets based on names of the anime shows.



DATA CLEANING

DROPPING UNNECESSARY COLUMNS

	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']

Step 1

There were some columns labeled “Unnamed” that were not needed in the Data Frame, so they were dropped to present a more accurate and concise data set.

```
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']

Step 2

DATA CLEANING

APPLYING STRING SPLITS TO DATA FRAME

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

	uid	Name	genre	aired	episodes	members	popularity	ranked	score	Type	Studio	Genres
0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shounen...']	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', 'Shounen...']	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', 'Fantasy', '...']	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']

```
In [ ]: df.drop('aired', axis = 1, inplace = True)
Out [ ]:
```

```
In [ ]: air_date_split = df['aired'].str.split(' to ')
df['Start Date'] = air_date_split.apply(lambda x : x[0])
df['End Date'] = air_date_split.apply(lambda x : x[1] if len(x) > 1 else None)
Out [ ]:
```

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

	uid	Name	genre	episodes	members	popularity	ranked	score	Type	Studio	Genres	Start Date	End Date
0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shounen...']	25.0	489888	141	25.0	8.82	TV	['Production I.G']	['Sports']	Oct 4, 2015	Mar 27, 2016
1	23273	Shigatsu wa Kimi no Uso	['Drama', 'Music', 'Romance', 'School', 'Shounen...']	22.0	995473	28	24.0	8.83	TV	['A-1 Pictures']	['Drama', 'Romance']	Oct 10, 2014	Mar 20, 2015
2	34599	Made in Abyss	['Sci-Fi', 'Adventure', 'Mystery', 'Drama', 'Fantasy', '...']	13.0	581663	98	23.0	8.83	TV	['Kinema Citrus']	['Adventure', 'Drama', 'Fantasy', 'Mystery', '...']	Jul 7, 2017	Sep 29, 2017
3	5114	Fullmetal Alchemist: Brotherhood	['Action', 'Military', 'Adventure', 'Comedy', '...']	64.0	1615084	4	1.0	9.23	TV	['Bones']	['Action', 'Adventure', 'Drama', 'Fantasy']	Apr 5, 2009	Jul 4, 2010
4	31758	Kizumonogatari III: Reiketsu-hen	['Action', 'Mystery', 'Supernatural', 'Vampire']	1.0	214621	502	22.0	8.83	Movie	['Shaft']	['Action', 'Mystery', 'Supernatural']	Jan 6, 2017	None

```
In 76: df = pd.read_csv('Cleaned_Main_Data (2).csv')
Out 76: df.head()
Out 76: df.head()
```

	uid	Name	genre	episodes	members	popularity	ranked
0	28891	Haikyuu!! Second Season	['Comedy', 'Sports', 'Drama', 'School', 'Shounen']	25.0	489888	141	
1	23273	Shigatsu wa Kimi no Uso	['Drama', 'Music', 'Romance', 'School', 'Shounen']	22.0	995473	28	
2	34599	Made in Abyss	['Sci-Fi', 'Adventure', 'Mystery', 'Drama', 'Fantasy']	13.0	581663	98	
3	5114	Fullmetal Alchemist: Brotherhood	['Action', 'Military', 'Adventure', 'Comedy', 'Drama', 'Magic...']	64.0	1615084	4	
4	31758	Kizumonogatari III: Reiketsu-hen	['Action', 'Mystery', 'Supernatural', 'Vampire']	1.0	214621	502	

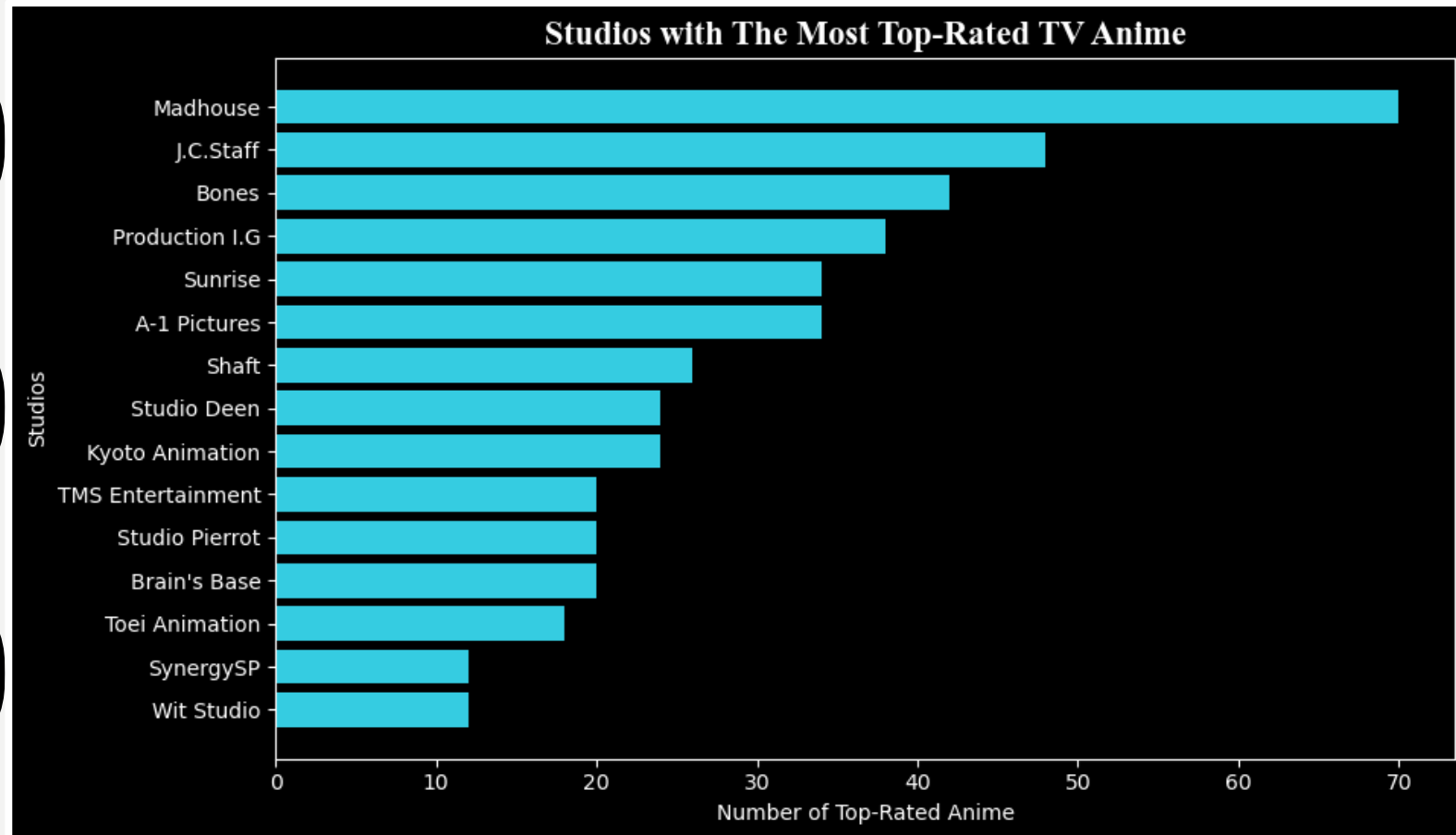
```
In 77: genre_split = df['genre'].str.split(',')
df['Genre'] = genre_split.apply(lambda x : x[0])
Out 77: df.head()
Out 77: df.head()
```

	episodes	members	popularity	ranked	score	Type	Studio	Start Date	End Date	Genre
0	25.0	489888	141	25.0	8.82	TV	['Production I.G']	Oct 4, 2015	Mar 27, 2016	['Comedy']
1	22.0	995473	28	24.0	8.83	TV	['A-1 Pictures']	Oct 10, 2014	Mar 20, 2015	['Drama']
2	13.0	581663	98	23.0	8.83	TV	['Kinema Citrus']	Jul 7, 2017	Sep 29, 2017	['Sci-Fi']
3	64.0	1615084	4	1.0	9.23	TV	['Bones']	Apr 5, 2009	Jul 4, 2010	['Action']
4	1.0	214621	502	22.0	8.83	Movie	['Shaft']	Jan 6, 2017	None	['Action']

We had to split strings in our "aired" column to create two new columns to represent a Start Date & End Date. We performed a split for genre as well. Our genre column had concatenated strings, so we had to run a code in which each row in genre would only have the first classification listed. We believe this was best to convey the true (or dominant) category of a show.

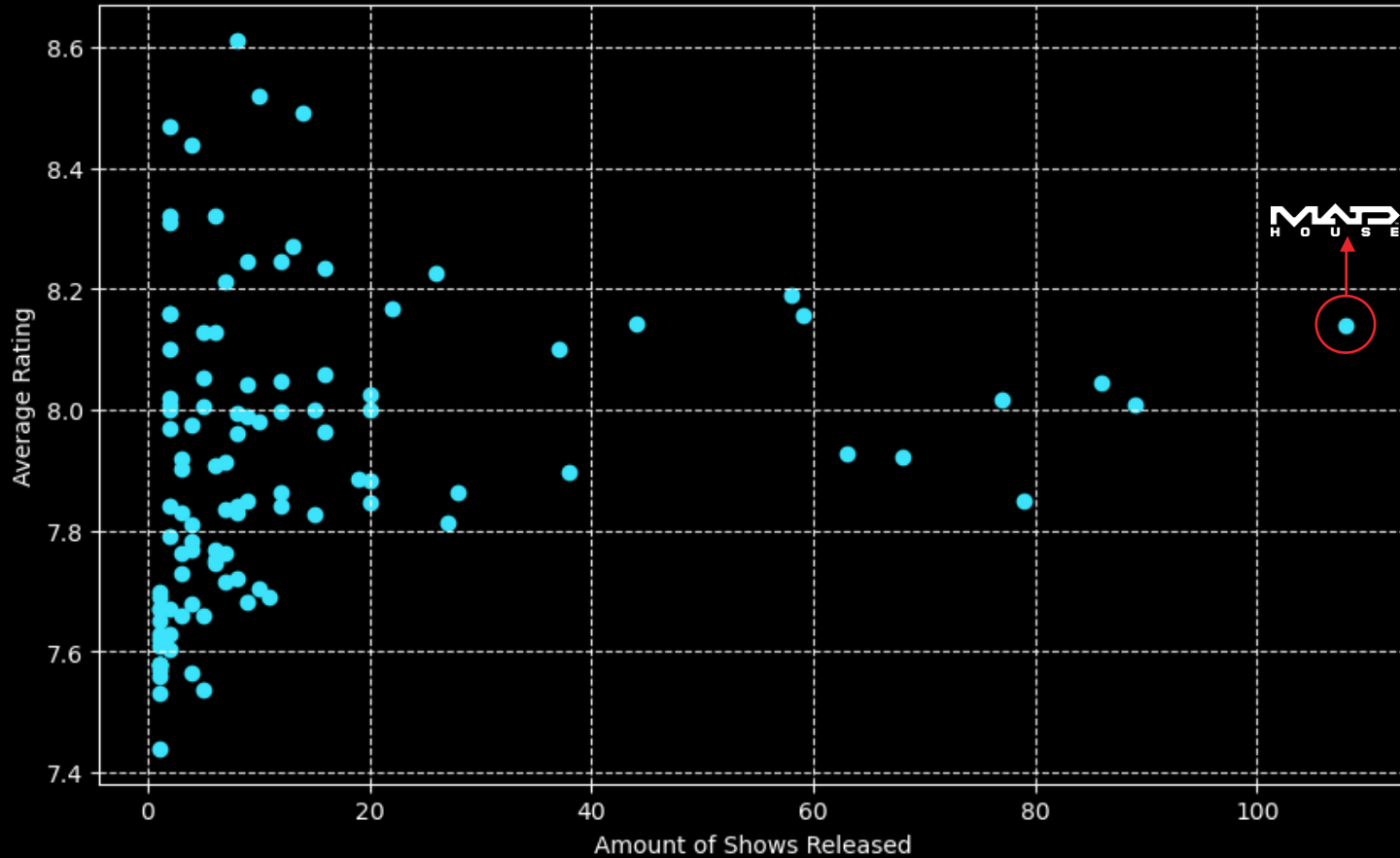
QUESTION 1

ANIME STUDIO WITH THE MOST TOP-RATED ANIME



The animation studio, Madhouse, is shown to have the most top-rated anime, with a total of 70 shows. This studio has a total of 70 anime shows that have a rating of 8 or higher. They have produced numerous popular anime you may have heard of, such as HunterXHunter, Death Note, and One Punch Man.

Average Rating by Amount of Anime Shows

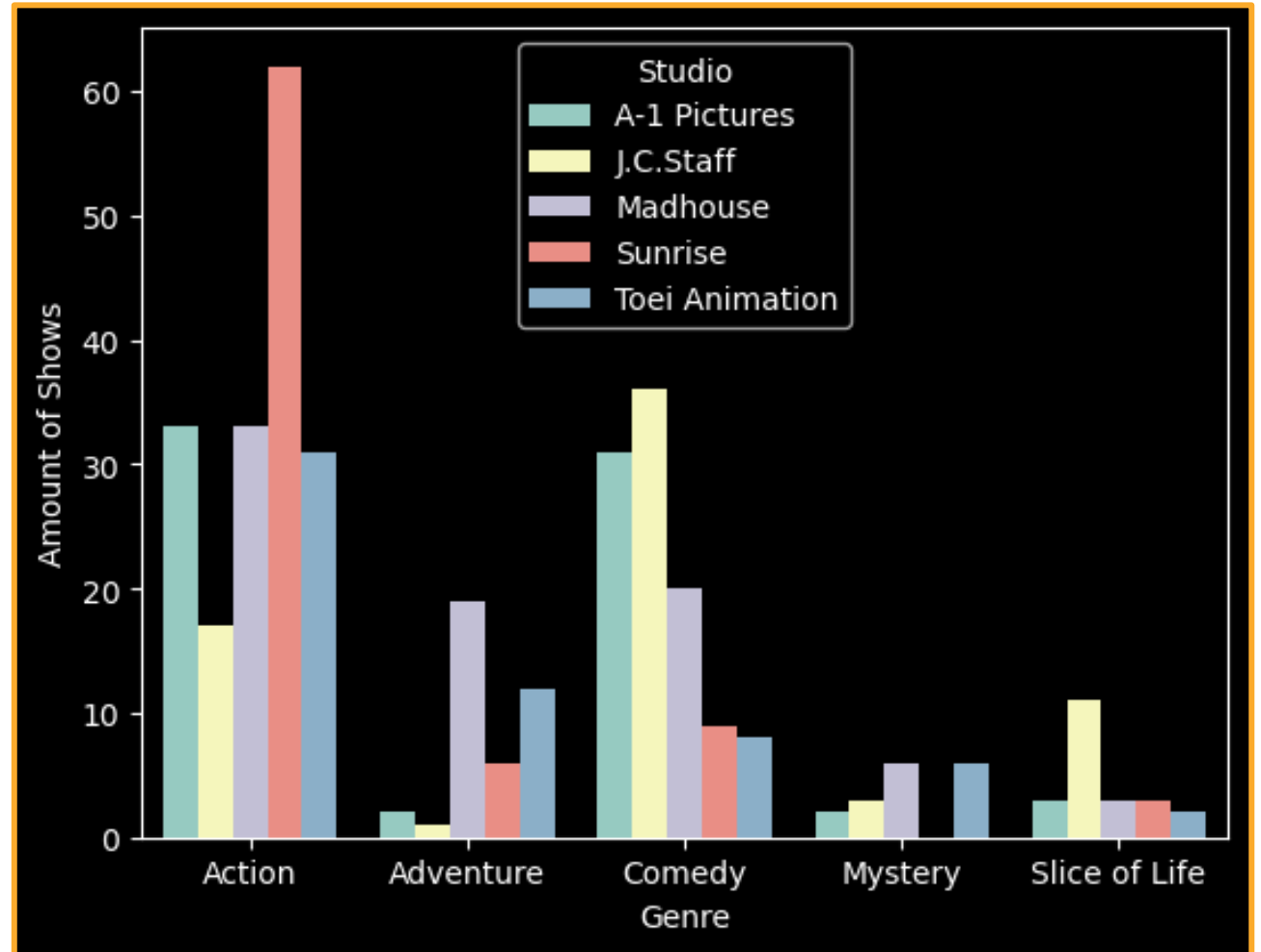


WOW ! HOW GREAT IS THAT ?

Seeing that most of the studios have produced 20 anime shows or less, with their rating averaging from 7 to 8, this shows how one of a kind Madhouse Studio truly is. They have produced over 108 anime and yet have an average rating of 8. Did we mention this studio has been around since 1972?! For a studio to stay this consistent for this amount of time is a rare feat.



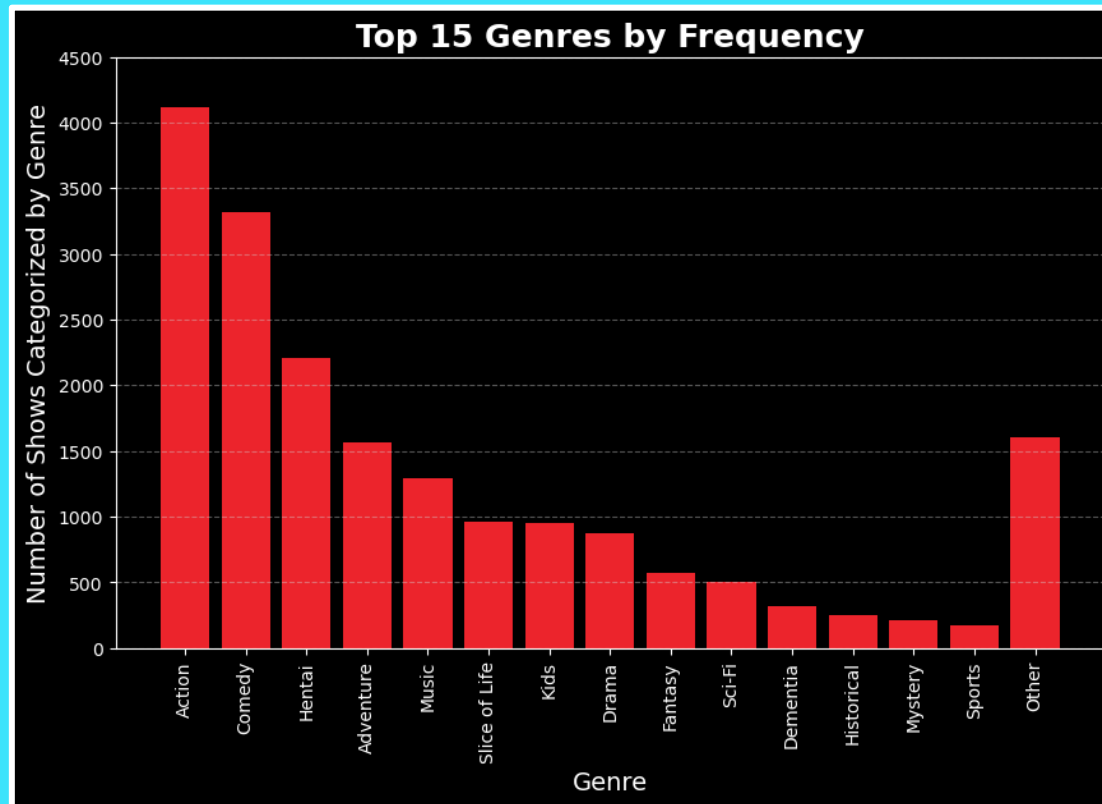
NUMBER OF SHOWS BY GENRE



This grouped bar chart represents the shows by genre from the top 5 studios. As you can see, although Madhouse may dominate the rating charts, they do not stick to one specific genre, while others, such as Sunrise, appear to focus in on a preferred genre. This could possibly be why Madhouse has success due to how versatile they are.

QUESTION 2: WHAT ARE THE MOST POPULAR GENRES IN ANIME?

- Just as colorful as the shows can be, same is the spectrum of genres associated with anime. Many of which you are probably familiar with, however, some may be higher on the list than one may expect. As you can see, a vast majority fall into the Action category with Comedy falling close behind.



For Consideration: Full Genre List
By Quantity of Appearances in Data Set

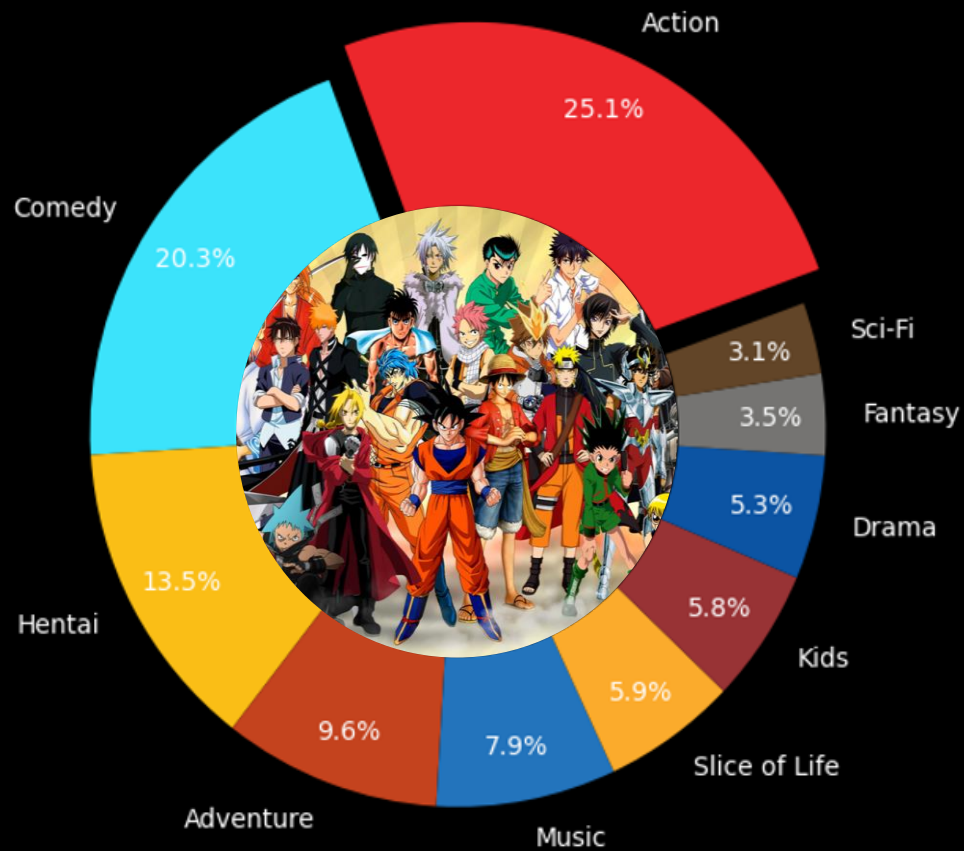
Action	4120	Game	176
Comedy	3325	Romance	158
Hentai	2212	Harem	156
Adventure	1567	Ecchi	146
Music	1296	Magic	112
Slice of Life	963	Military	108
Kids	948	Mecha	82
Drama	876	Demons	82
Fantasy	573	Horror	78
Sci-Fi	504	Parody	78
Dementia	317	Cars	60
Historical	247	Supernatural	56
Mystery	209	School	52
Sports	177	Psychological	49
		Space	36
		Shounen	21
		Police	14
		Super Power	11
		Seinen	11

↑ (Top 14)

(Remainder with
quantity above 10) ↑

Are any of these genres your favorite?
Are there any left out? Any you would include?

Top 10 Anime Genres



HMMM...

With two categories making up most of the genre count, where does this leave the others? There is a steep difference between those at the top of the list and those at the tail end. These niche few still hold their place in the list as many were listed as 'sub-categories' in the original data sets. All genres help shape the viewing landscape with diversity and depth. Given the pure nature of anime, we originally assumed genres, such as Fantasy and Super Power, would sit higher on the pedestal.

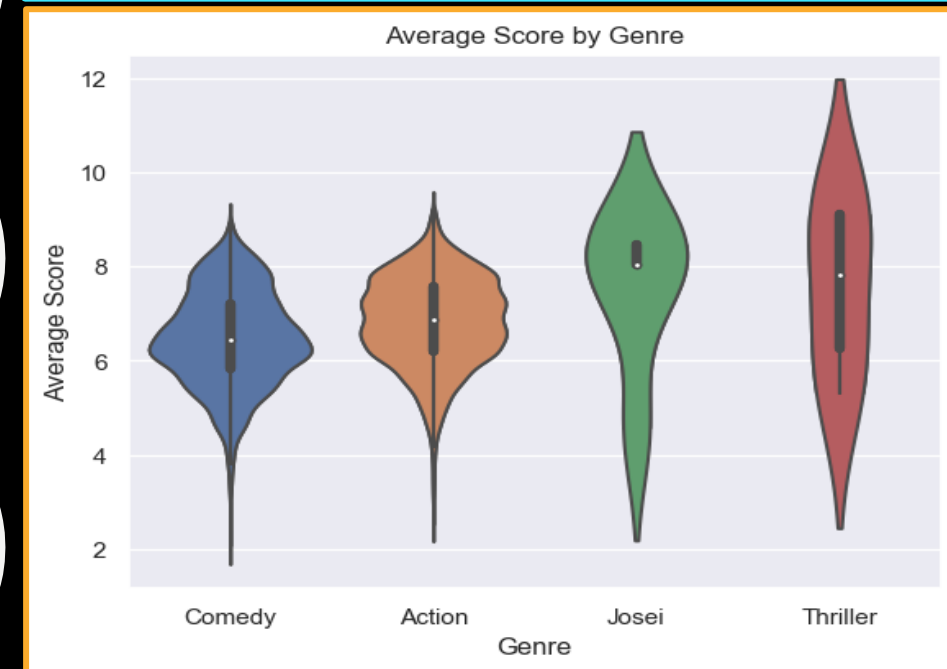
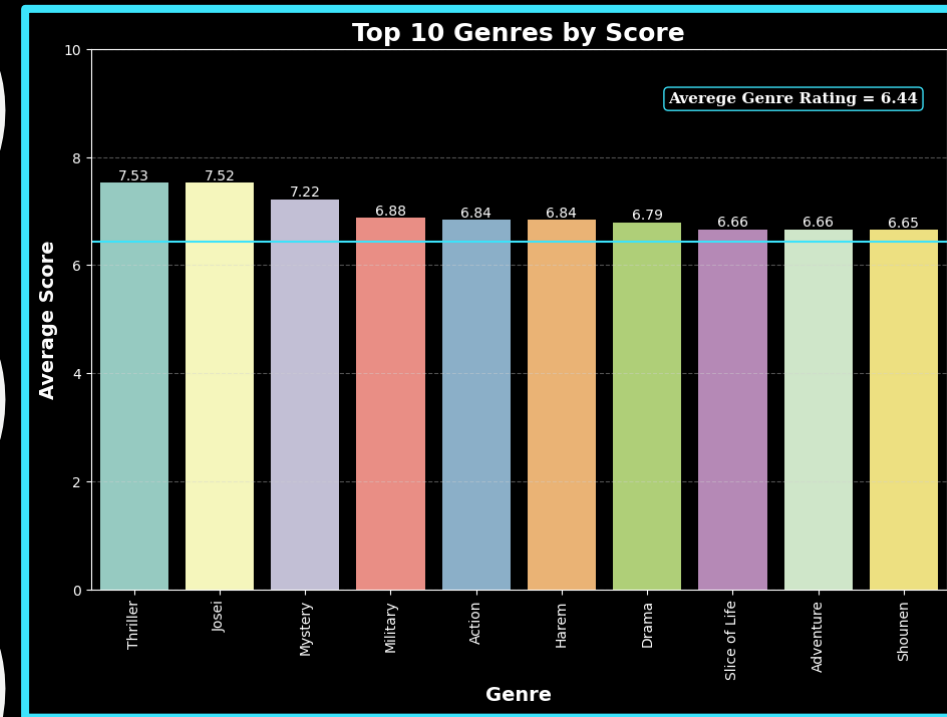
Some genres we did not necessarily expect to be most common were Music and Slice of Life. This assumes there is a large fan base for slower, more casual viewers. Even as avid anime watchers ourselves, these are not ones we personally tend to lean towards, proving there is a show with a classification and plot for everyone.

HOWEVER...

If most of the anime that gets produced is categorized under Action, can we assume this means it is also the best rated?

As shown by the bar graph, Action is ranked 5th while Thriller takes first place with an average score of 7.53.

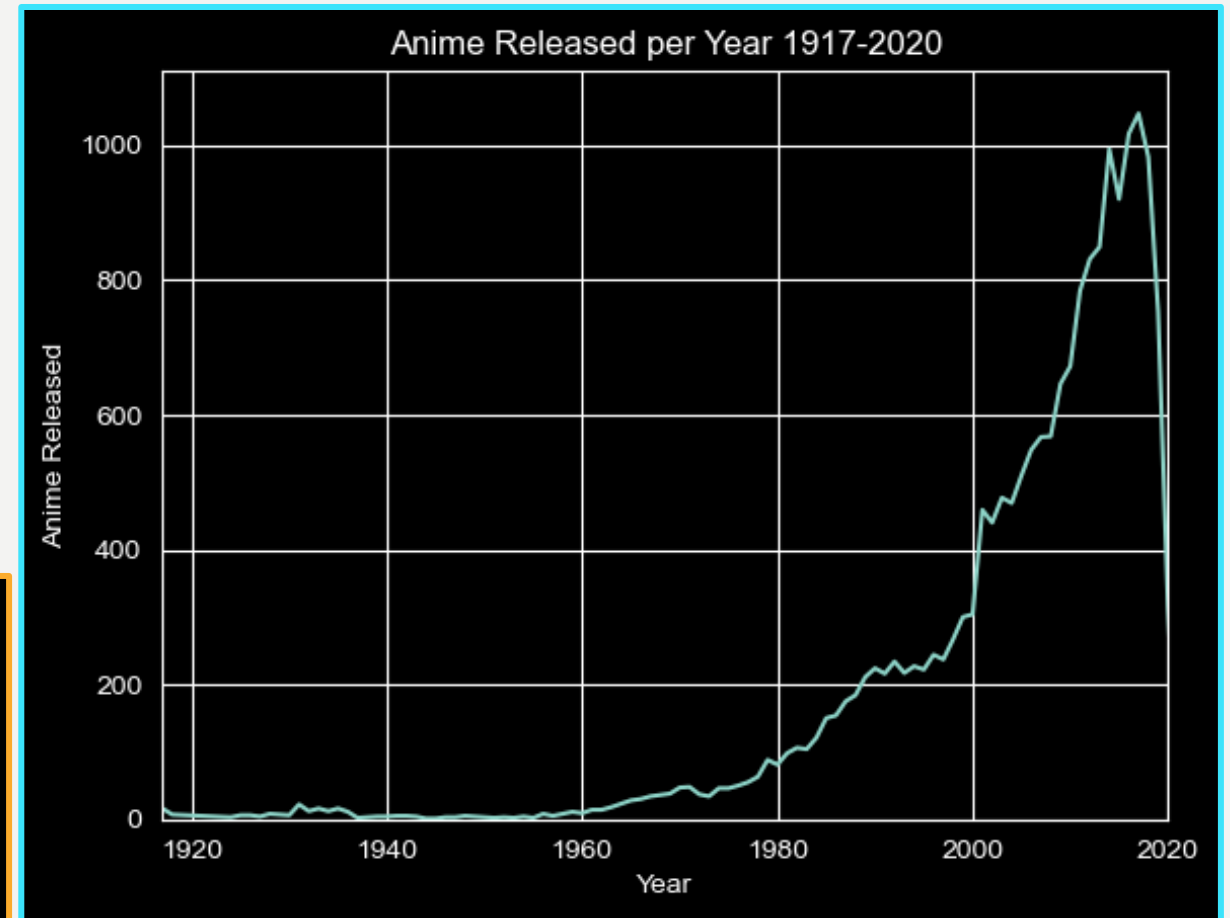
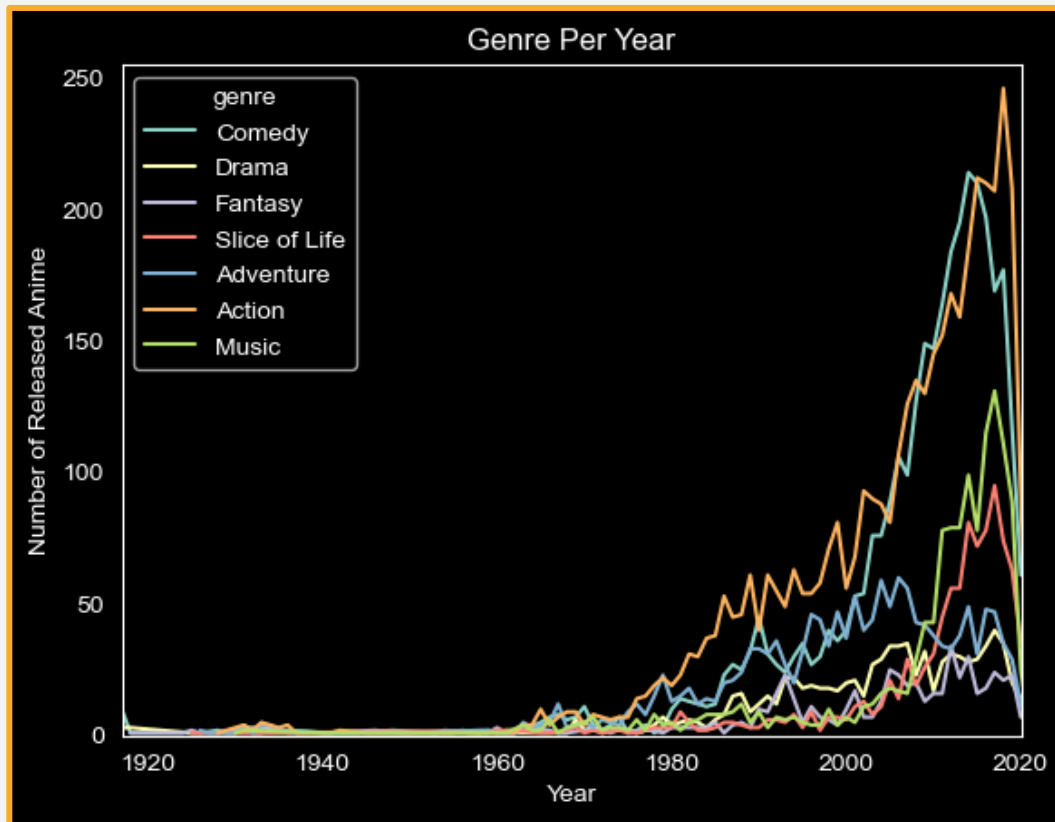
With help of a violin plot we can compare the 2 top-rated genres to the 2 most-common genres. As you will see, both Comedy and Action experienced some outliers possibly affecting median score, while Thriller and Josei remained relatively consistent with their score votes.



QUESTION 3

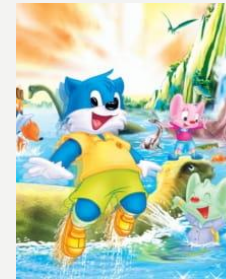
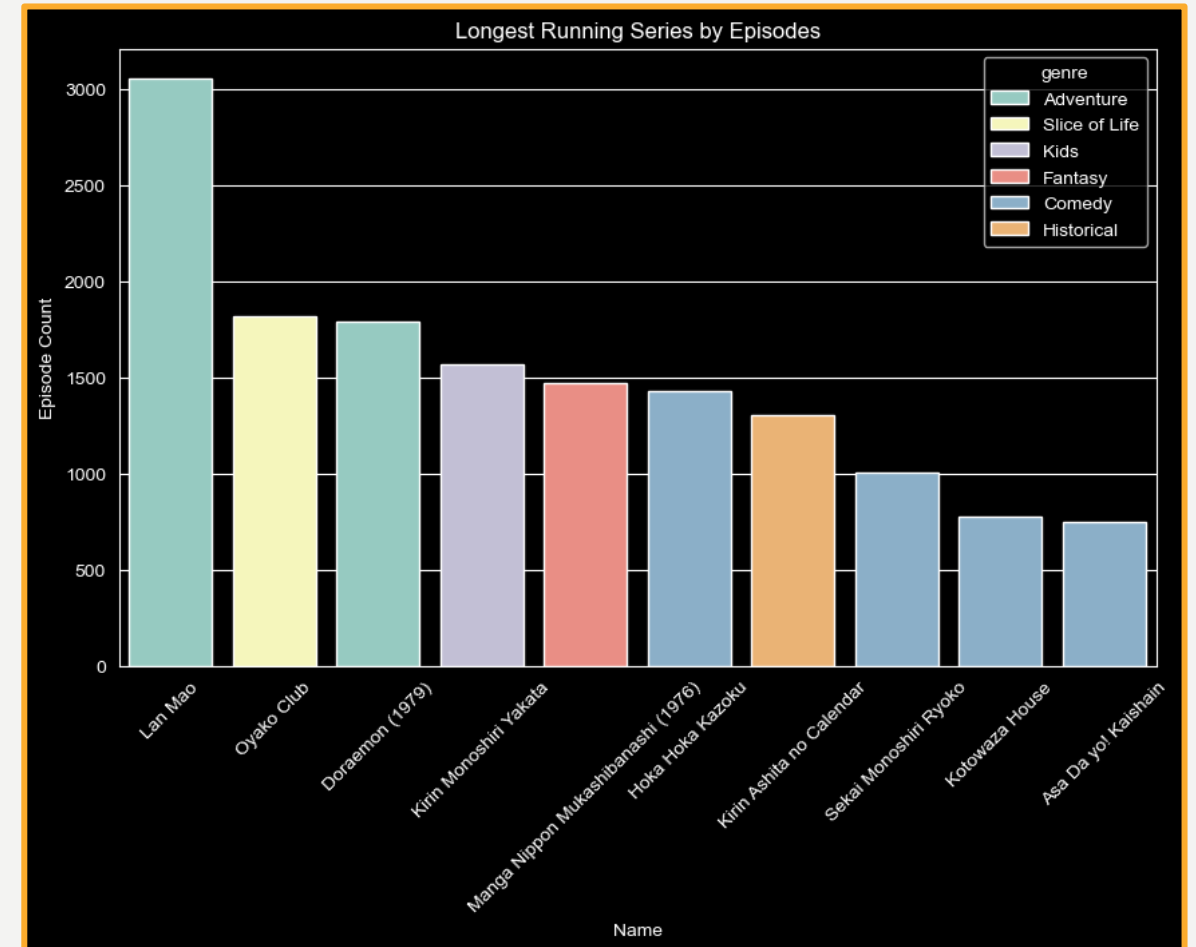
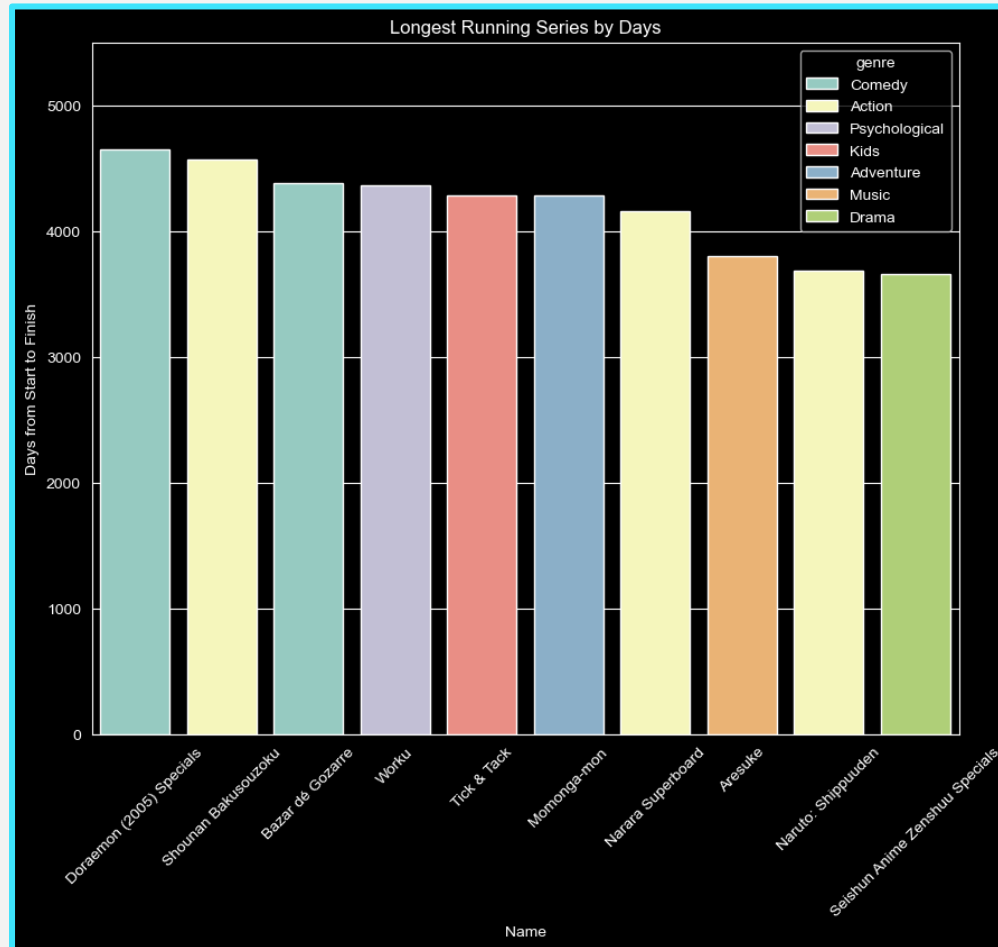
LOOKING AT THE GROWTH OF ANIME

The growth of anime has been explosive across all genres, but what genres have grown the most rapidly?



There is a possibility our dataset is incomplete, however I am inclined to believe this massive drop off we see in 2020 is accurate within an acceptable margin, the cause? The extreme measures taken by the Japanese government to curb coronavirus and an inability by the industry at large to adjust.

The longest running series have only been able to go on so long because they have been able to consistently find funding for productions, let's look at trends among them.



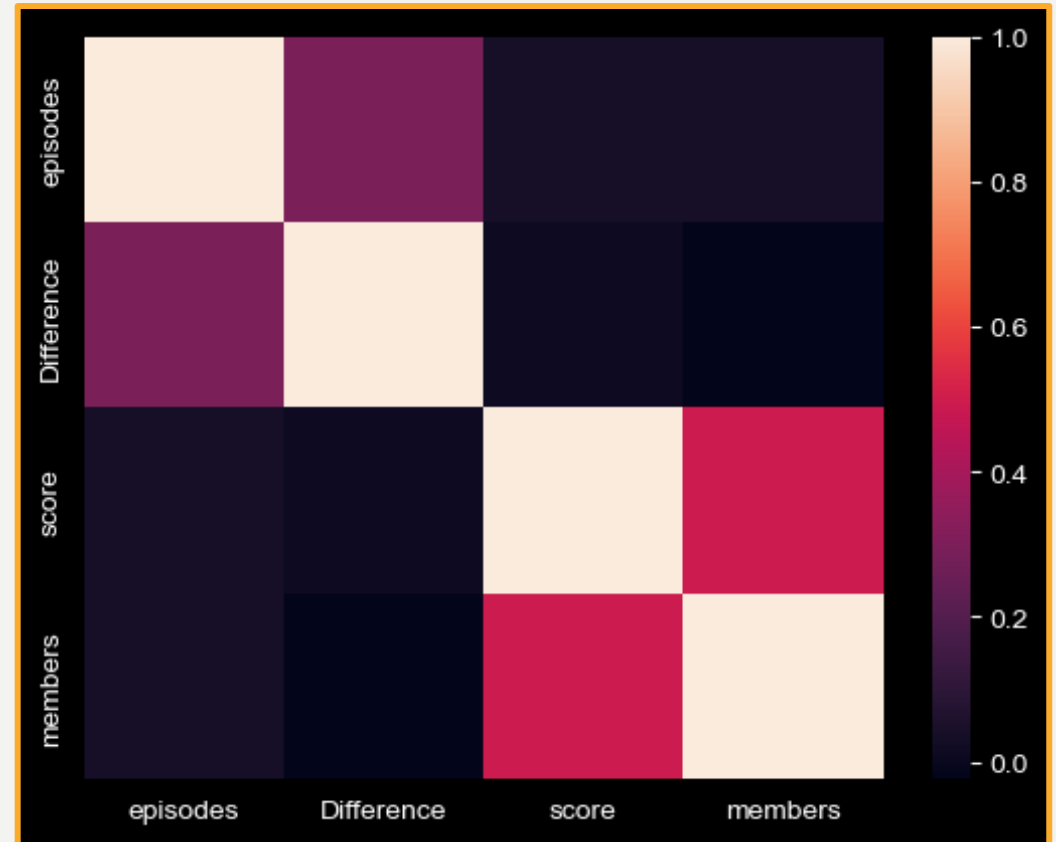
LINEAR REGRESSION

OR HOW I LEARNED TO STOP WORRYING AND LOVE THE NUMBERS

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, it is not great. However, for our purposes it will be a fine representation.



	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



```
X1 = df11.loc[:,['Difference','episodes','members']]
y1 = df11.score
```

```
X1 = sm.add_constant(X1)
```

```
model = sm.OLS(y1, X1).fit()
predictions1 = model.predict(X1)
```

```
print_model = model.summary()
print(print_model)
```

Executed in 45ms, 17 May at 14:52:34

```

=====
                        OLS Regression Results
=====
Dep. Variable:          score    R-squared:          0.191
Model:                  OLS      Adj. R-squared:      0.190
Method:                 Least Squares    F-statistic:    136.6
Date:                   Wed, 17 May 2023    Prob (F-statistic): 1.75e-79
Time:                   14:52:34    Log-Likelihood:   -408.84
No. Observations:      1741    AIC:                825.7
Df Residuals:          1737    BIC:                847.5
Df Model:               3
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	7.8579	0.012	664.715	0.000	7.835	7.881
Difference	-1.186e-05	4.46e-05	-0.266	0.790	-9.94e-05	7.57e-05
episodes	0.0004	0.000	1.109	0.268	-0.000	0.001
members	5.582e-07	2.81e-08	19.881	0.000	5.03e-07	6.13e-07

```

=====
Omnibus:                 100.611    Durbin-Watson:      0.649
Prob(Omnibus):            0.000    Jarque-Bera (JB):    127.856
Skew:                     0.547    Prob(JB):            1.72e-28
Kurtosis:                 3.751    Cond. No.            5.48e+05
=====

```

The resultant R-squared values are extremely low for both the multiple linear regression (Left) and the simple linear regression (Right). Both of which provided near identical values of 0.191 and 0.19.

```
X2 = df11.loc[:,['members']]
y2 = df11.score
```

```
X2 = sm.add_constant(X2)
```

```
model = sm.OLS(y2, X2).fit()
predictions2 = model.predict(X2)
```

```
print_model = model.summary()
print(print_model)
```

Executed in 31ms, 17 May at 14:52:44

```

=====
                        OLS Regression Results
=====
Dep. Variable:          score    R-squared:          0.190
Model:                  OLS      Adj. R-squared:      0.190
Method:                 Least Squares    F-statistic:    408.1
Date:                   Wed, 17 May 2023    Prob (F-statistic): 1.11e-81
Time:                   14:52:44    Log-Likelihood:   -409.80
No. Observations:      1741    AIC:                823.6
Df Residuals:          1739    BIC:                834.5
Df Model:               1
Covariance Type:       nonrobust
=====

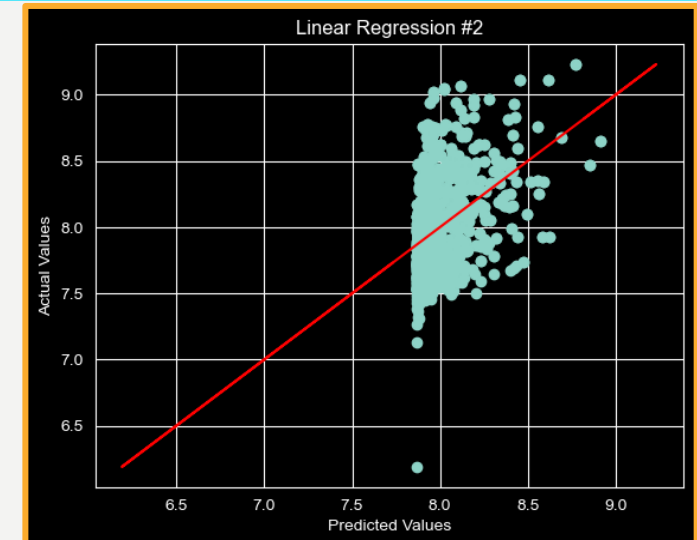
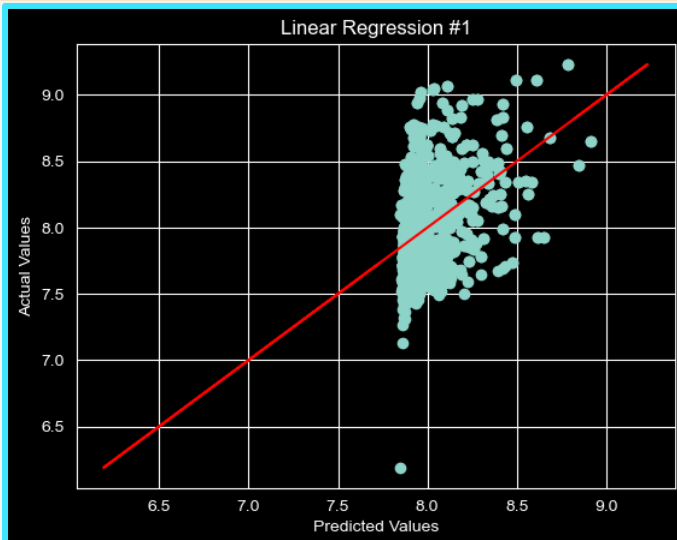
```

	coef	std err	t	P> t	[0.025	0.975]
const	7.8653	0.009	831.940	0.000	7.847	7.884
members	5.611e-07	2.78e-08	20.200	0.000	5.07e-07	6.16e-07

```

=====
Omnibus:                 102.400    Durbin-Watson:      0.650
Prob(Omnibus):            0.000    Jarque-Bera (JB):    131.380
Skew:                     0.550    Prob(JB):            2.96e-29
Kurtosis:                 3.775    Cond. No.            4.38e+05
=====

```



STATISTICAL HYPOTHESIS

Given that anime is released on a seasonal basis regardless of genre, I hypothesized that all genres would share a similar airing span. To check my hypothesis I created a 'difference' value which represents the days between the first airing date of a series and the respective finale, where possible*, and compared the averages across genres.



```
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)
```

However, since we can see that the p-value of our t-tests are 0.02 and 0.03, 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.

*-See limitations 3

CONCLUSION



Thriller –

Erased – 8.3/10



Josei –

Yuri on Ice – 7.9/10



Mystery –

Monster – 8.9/10



Military –

Full Metal Alchemist Brotherhood - 9.1/10



Action -

HunterXHunter – 9.0/10



ANIME RECOMMENDATIONS

Haven't dived into the world of Anime?

Here are some good anime we recommend giving a try based on the top 5 genres with highest average scores.

LIMITATIONS & BIAS

WHAT KEPT US FROM GOING ULTRA INSTINCT

- 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. I felt like 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 interpolate another ranking system from a mainly 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. I was forced to drop a large 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.
- Similar to the previous limitation, series which took an extended hiatus where there was no new material for long periods of time 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. Thusly, there may be multiple series which 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 on 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.

- Which areas in the world have the highest density of viewers?
- Do certain locations have a preference of 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?

THANK YOU

ありがとう

ARIGATŌ

