ANIME AND USERS ANALYSIS

MyAnimeList.net

IST652 GROUP 7

PEILIN ZHONG, SIJIN ZHOU, ZEYI LUO

Table of Contents

Abstract

Otaku is a Japanese term describing those who are obsessed with some pop cultures, such as manga and anime. Although it originated as a negative stereotype, now there are an increasing number of people consuming interests in anime and manga and its meaning is becoming much less negative [1]. The otaku subculture continued to grow with the expansion of internet and social media. For now, the otaku subculture is one of the most famous Japanese pop cultures worldwide.

Project Goal

As one of those who interested in anime, we conduct this project with the initial aim of assisting animation studios to:

1. Identify seasonal trends of the anime market
2. Understand loyal users' watching and rating behavior
3. Conduct wise strategies for expanding their market shares in order to attract more target audience

Data Acquisition

We choose MyAnimeList.net as the data source for our study. MyAnimeList.net is one of the world's largest anime and manga database communities [2]. It allows users to create their own list of anime and share ideas in the community.

The data for this project is collected by web scraping from watching challenge 2015-2018 forum threads on MyAnimeList.net, which was half-prepared by Azathoth. We are going to analyze the works half-prepared by Azathoth. The web scraping data could be downloaded from Kaggle [3].

Our datasets are directly downloaded from Kaggle, there is few works left to do additional web scraping. In the part, we will research our problems by descriptive analysis. Basically, we will apply Python packages such as pandas, numpy and many others to summarize data to get insights about users' characteristics and the relationship between anime features and users' behaviors.

To efficiently identify patterns and trends, we will also demonstrate those problems by summarized structured tables as well as visualization.

Datasets Summary

There are three structured datasets used in this project.

The ‘anime_cleaned.csv’ dataset has 6,668 rows and 33 columns, containing features of anime like title, type, source, episodes, duration, rating, score, background, producer, studio, genre, aired_from_year etc.

The ‘users_cleaned.csv’ dataset has 108,711 rows and 17 columns, containing users’ information like username, status (watching, completed, on hold, dropped, plan to watch), gender, location, join date, last online etc.

The ‘animelists_cleaned.csv’ dataset has 31,284,030 rows and 11 columns, containing information of user-customized anime lists, like username, anime id, watched episodes, start / finish date, score, status, timestamp of last updated etc.

There are couple of interesting facts we find in the datasets listed above:

1. There are many extreme values and missing values in the datasets. Therefore, handling those extreme values and nulls will be the core of our data preprocessing.
2. In the dataset ‘anime_cleaned.csv’, there are multiple values in each single cell of the ‘genre’ column. Therefore, we also need to split then aggregate the anime genre properly for further analysis.
3. In the dataset ‘anime_cleaned.csv’, there are some categorical variables that have too specific / messy categories, thus reclassifying / normalizing those sub-categories should be taken into consideration.
4. In the dataset ‘anime_cleaned.csv’, there are multiple columns related to the performance of an anime, like score, rank, popularity. Therefore, clarifying the definition of each variable is essential.

Information of ‘anime_cleaned.csv’:

```
0   anime_id         6668 non-null   int64
1   title            6668 non-null   object
2   title_english    3438 non-null   object
3   title_japanese   6663 non-null   object
4   title_synonyms   4481 non-null   object
5   image_url        6666 non-null   object
6   type             6668 non-null   object
7   source           6668 non-null   object
8   episodes         6668 non-null   int64
9   status           6668 non-null   object
10  airing           6668 non-null   bool
11  aired_string     6668 non-null   object
12  aired            6668 non-null   object
13  duration         6668 non-null   object
14  rating           6668 non-null   object
15  score            6668 non-null   float64
16  scored_by        6668 non-null   int64
17  rank             6312 non-null   float64
18  popularity       6668 non-null   int64
19  members          6668 non-null   int64
20  favorites        6668 non-null   int64
21  background       813 non-null    object
22  premiered        2966 non-null   object
23  broadcast        2980 non-null   object
24  related          6668 non-null   object
25  producer         4402 non-null   object
26  licensor         2787 non-null   object
27  studio           6668 non-null   object
28  genre            6664 non-null   object
29  opening_theme    6668 non-null   object
30  ending_theme     6668 non-null   object
31  duration_min     6668 non-null   float64
32  aired_from_year  6668 non-null   float64
```

Information of ‘Users_Cleaned.csv’:

```
0   username                  108710 non-null  object
1   user_id                   108711 non-null  int64
2   user_watching             108711 non-null  int64
3   user_completed            108711 non-null  int64
4   user_onhold               108711 non-null  int64
5   user_dropped              108711 non-null  int64
6   user_plantowatch          108711 non-null  int64
7   user_days_spent_watching  108711 non-null  float64
8   gender                    108711 non-null  object
9   location                  108706 non-null  object
10  birth_date                108711 non-null  object
11  access_rank               0 non-null       float64
12  join_date                 108711 non-null  object
13  last_online               108711 non-null  object
14  stats_mean_score          108711 non-null  float64
15  stats_rewatched           108711 non-null  float64
16  stats_episodes            108711 non-null  int64
dtypes: float64(4), int64(7), object(6)
```

Information of ‘animelists_cleaned.csv’:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31284030 entries, 0 to 31284029
Data columns (total 11 columns):
 #   Column               Dtype  
---  ------               -----  
 0   username             object 
 1   anime_id             int64  
 2   my_watched_episodes  int64  
 3   my_start_date        object 
 4   my_finish_date       object 
 5   my_score             int64  
 6   my_status            int64  
 7   my_rewatching        float64
 8   my_rewatching_ep     int64  
 9   my_last_updated      object 
 10  my_tags              object 
dtypes: float64(1), int64(5), object(5)
memory usage: 2.6+ GB
```

Data Preprocessing

According to the preliminary exploration, there are many unformatted, outliers or Nan values in these three datasets, thus it is essential to conduct thorough data cleaning and formatting to get workable datasets. The general procedures include - format data types, handle extreme values, handle missing values, examine duplicates, classification and subtraction.

Format Data Types

In order to give the appropriate outcome of describe() function, there are some data types formatting steps to take.

1. Anime

   Convert integer to string, such as 'anime_id'

   Convert floating number to string, such as 'aired_from_year'

2. User

   Convert integer to string, such as 'user_id'

   Convert string to datetime, such as 'birth_date', 'join_date', 'last_online'

3. Animelist

   Convert integer to string, such as 'anime_id'

Convert string to datetime, such as ‘my_last_updated’

Handle Extreme Values

Extreme values include extremely small and large values, which may add noise to the overall data distribution and lead to some insignificant analysis outcome. Therefore, we examine the extreme values for each numerical variable in these three datasets and apply appropriate methods to handle those extreme values. Throughout the describe() function, there are no negative values in the datasets, thus no specific processing on that.

```
anime.describe()
```

	episodes	score	scored_by	rank	popularity	members	favorites	duration_min
count	6668.000000	6668.000000	6.668000e+03	6312.000000	6668.000000	6.668000e+03	6668.000000	6668.000000
mean	14.276395	6.848998	2.403501e+04	4327.645120	4479.515897	4.749037e+04	670.365627	28.442167
std	40.906929	0.927448	6.112103e+04	3170.699074	3453.338080	1.051211e+05	3823.072834	25.365980
min	0.000000	0.000000	0.000000e+00	1.000000	1.000000	1.800000e+01	0.000000	0.000000
25%	1.000000	6.350000	6.812500e+02	1710.750000	1691.750000	2.222750e+03	3.000000	17.000000
50%	6.000000	6.930000	3.966000e+03	3754.500000	3629.500000	1.033650e+04	21.000000	24.000000
75%	13.000000	7.460000	1.976075e+04	6338.500000	6630.250000	4.336000e+04	142.000000	27.000000
max	1818.000000	9.520000	1.009477e+06	12856.000000	14468.000000	1.456378e+06	106895.000000	163.000000

```
user.describe()
```

	user_watching	user_completed	user_onhold	user_dropped	user_plantowatch	user_days_spent_watching	access_rank	stats_mean_score	stats_rewat
count	108711.000000	108711.000000	108711.000000	108711.000000	108711.000000	108711.000000	0.0	108711.000000	108711.00
mean	14.767503	196.458178	11.388167	11.733716	75.578589	61.913873	NaN	7.747612	14.19
std	32.746591	244.945751	30.830825	30.978991	178.653664	59.211762	NaN	1.451368	55.37
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.500694	NaN	0.000000	0.00
25%	3.000000	50.000000	0.000000	0.000000	6.000000	21.066319	NaN	7.330000	0.00
50%	7.000000	123.000000	4.000000	3.000000	27.000000	46.190278	NaN	7.890000	1.00
75%	16.000000	254.000000	12.000000	12.000000	81.000000	84.461806	NaN	8.460000	10.00
max	2934.000000	5479.000000	2562.000000	2457.000000	12051.000000	952.654595	NaN	10.000000	9404.00

```
animelists.describe()
```

	my_watched_episodes	my_score	my_status	my_rewatching	my_rewatching_ep
count	3.128403e+07	3.128403e+07	3.128403e+07	2.440578e+07	3.128403e+07
mean	1.289615e+01	4.652315e+00	3.008111e+00	7.903045e-04	1.832994e-01
std	3.733380e+01	3.931615e+00	1.730421e+00	2.810124e-02	1.009172e+03
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00	2.000000e+00	0.000000e+00	0.000000e+00
50%	4.000000e+00	6.000000e+00	2.000000e+00	0.000000e+00	0.000000e+00
75%	1.300000e+01	8.000000e+00	4.000000e+00	0.000000e+00	0.000000e+00
max	9.999000e+03	1.000000e+01	5.500000e+01	1.000000e+00	5.644513e+06

0 Values

The common method for handling extreme values is keep, replace or drop. In general, we keep those that have specific meaning; replace those that are resulting from entry error; drop those that have no alternative for replacement, nor no necessity to keep. Drop would be the very last choice, since we want to maintain the data variance of the cleaned datasets as those before preprocessing.

1. **anime**

   We keep those 6 records in the 'score' column where values equal to 0, because no one has scored that anime yet (the corresponding values in 'score_by' column are also 0). Similarly, we keep those 851 records in 'favorites' column where values equal to 0, because no one has given a favorite to those anime yet.

   There are 133 records where 'episodes' are 0, 47 records where 'duration_min' are 0. These 180 records are dropped, because there is not sufficient alternative information to support the replacement.

2. **user**

   Technically, it makes sense for variables like user_watching, user_completed, user_onhold, user_dropped, user_plantowatch, stats_mean_score, stats_rewatched to have 0 values. Therefore, no replacement or drop needed for 'user' dataset.

3. **animelists**

   Technically, it makes sense for these variables (my_watched_episodes, my_score, my_rewatching) to have 0 values. Therefore, no imputation or drop needed.

   For dummy variables, we treat differently. 'my_rewatching' is a dummy variable (0: no rewatching, 1: rewatching). We cross analysis 'my_rewatching' column together with 'my_rewatching_ep' column - When my_rewatching_ep > 0, we replace 'my_rewatching' with 1 (4285 rows affected); When my_rewatching_ep = 0: keep the same, since a user may start to rewatch a anime but have not finished the entire episode, which would not be counted as a completed rewatching episode in this case (18789 rows match this situation).

   'my_start_date', 'my_finish_date': There are '0000-00-00' values (84.45% of the dataset) in these two columns, which do not give any useful information, thus these to columns will not be our focused variables in the following analysis

Extremely Large Values

1. **anime**

Based on common sense, TV series animes have way more episodes than any other type.

Therefore, we check on TV anime whose episodes are extremely large to explore the underlying details. There are 90 TV animes with more than 100 episodes; 11 TV animes with more than 100 episodes; only 3 TV animes with more than 100 episodes. These tremendous works are Ninja Hattori-kun, Doraemon and Oyako Club.

2. **user, animelists**

   For those users whose anime status statistics are extremely high, we attempt to apply cross-check on user_watching, user_completed, user_days_spent_watching, join_date, in order to find out whether there are nonsense records in each column, like a newly-joined user has abnormally number of completed anime.

   However, the workload for that is extremely and not practical, since most of the user info is customized and could be modified by users. There are thousands of nonsense self-defined location info. Besides, we also blamed the web scraping accuracy. Based on the sampling check, many users update their info after the scraping period, thus the majorities are out of date. Therefore, we need to clarify the analysis duration for these datasets, and also hold high tolerance for generating conclusions based on unqualified data.

   Same issues for animelists dataset.

Handle Missing Values

For each dataset, we first summarize the number of rows with NA values for each column in the dataframe; then deep-dive into each one of those columns with NA values, together with cross analysis of other columns in the same dataframe, in order to determine which type of those NA values are and what should be the correspondingly proper method to fix them.

1. **anime**

   There are 356 records where ‘rank’ are NA values, and 355 of them have ‘rating’ as R - 17+ (violence & profanity) / R+ - Mild Nudity / Rx - Hentai. According to the ranking cretiarias [4] on MyAnimeList.net, any anime with R related rating is excluded from the rankings, thus those anime will not have a rank. These NA values are intended to be left blank, thus no additional action should be taken. Only 1 anime is rated as 'PG-13 - Teens 13 or older', thus, only replace this one record with the 'rank' value of the anime with the same score.

```
anime[(anime['rank'].isnull()) & (anime['rating'] == 'PG-13 - Teens 13 or older')].score
# anime_id = 6546, score = 5.07

# Replace the NA value in 'rank' with the 'rank' value of the anime with same score
anime['rank'] = anime['rank'].mask(anime['score'] == 5.07, 8977)
```

There are 4 records where ‘genre’ is null. Because of the limited cases, we just research on the MyAnimeList.net website then manually replace those missing values with updated genre by indexing. 2 of those 4 records are updated, the remaining 2 records are dropped since there is not sufficient information for the replacement.

```
# Update the genre of anime_id = 17813 and anime_id = 37018
anime['genre'][5111] = 'Supernatural'
anime['genre'][6642] = 'Kids'

# anime_id = 33389 and anime_id = 32695, these two still don't have any genre set yet
# Drop these two records
anime = anime.drop(index = 2357).drop(index = 3301).reset_index(drop=True)
```

2. **user**

There is 1 record where ‘username’ is null, and no alternative way to replace it. It is dropped.

The entire ‘access_rank’ column is empty, thus we drop this column in the following subtraction section.

3. **animelists**

There are 243 records where ‘username’ is null, and no alternative way to replace it. These records are dropped.

There are 6,878,247 records where ‘my_rewatching’ is null. 206 of those values in ‘my_rewatching’ column are null while their corresponding 'my_rewatching_ep' != 0, which does not make any sense, thus we drop these 206 records. For the remaining 6,878,009 records where ‘my_rewatching’ is null and 'my_rewatching_ep' == 0, we fill those missing values with 0.

```
animelists[(animelists['my_rewatching'].isnull()) & (animelists['my_rewatching_ep'] != 0)]
# 206 rows in 'my_rewatching' is null while 'my_rewatching_ep' != 0,
# which does not make much sense, thus drop these NA values

# Drop all rows where 'my_rewatching' is null while 'my_rewatching_ep' != 0
my_rewatching_drop_index = animelists[(animelists['my_rewatching'].isnull()) & (animelists['my_rewatching_ep'] != 0)].i
animelists = animelists.drop(my_rewatching_drop_index, axis = 'index').reset_index(drop = True)
```

```
animelists[(animelists['my_rewatching'].isnull()) & (animelists['my_rewatching_ep'] == 0)]
# 6878009 rows in 'my_rewatching' is null and 'my_rewatching_ep' = 0, which makes sense
# Fill out these NAs with 0
```

Examine Duplicates

There are no duplicate values in all of the three datasets.

Transformation and Classification

Extract the aired starting month from the 'aired_string' column, and add it as a new column of the 'anime' dataframe named 'aired_from_month', data type: string.

Categorize 'aired_stat_month' by season, create a new column called 'premiered_season', with values as 'Spring', 'Summer', 'Autumn' and 'Winter', data type: string.

```
# Extract aired starting month from column 'aired_string' in user dataset
aired_from_month = []
for x in anime.aired_string:
    aired_from_month.append(x[0:3])

# Add it as a new column to dataframe 'anime'
anime['aired_from_month'] = aired_from_month
```

```
# Classify quarters based on 'aired_from_month'
# Q1 / Spring: January, February, March
# Q2 / Summer: April, May, June
# Q3 / Autumn: July, August, September
# Q4 / Winter: October, November, December

Q1 = ['Jan', 'Feb', 'Mar']
Q2 = ['Apr', 'May', 'Jun']
Q3 = ['Jul', 'Aug', 'Sep']
Q4 = ['Oct', 'Nov', 'Dec']

# Create a new column 'premiered_season' and add it to dataframe 'anime'
premiered_season = []
for x in anime['aired_from_month']:
    if x in Q1:
        premiered_season.append('Spring')
    elif x in Q2:
        premiered_season.append('Summer')
    elif x in Q3:
        premiered_season.append('Autumn')
    elif x in Q4:
        premiered_season.append('Winter')
    else:
        premiered_season.append('Not Clear')

# Add it as a new column to dataframe 'anime'
anime['premiered_season'] = premiered_season
```

Reclassify some categorical variables in order to organize the categories for better summary and visualization. In general, we combine some subtypes based on their definition into a parent category. We implement this reclassification / normalization on 'source', 'rating' and 'genre' columns in the 'anime' dataframe.

For the 'source' column, we combine Manga, 4-koma manga, Web manga, Digital manga as one distinct source as 'Manga'; combine Light novel, Visual novel, Novel, Picture Book, Book as one distinct source as 'Novel (Book)'; combine Game, Card Game as one distinct source as 'Game'; assign the remaining subtypes as source 'Other'. There are 5 unique source categories after the reclassification.

```python
# Re-classify / Normalize the 'source' in anime_sub
# Combine some subtypes into a parent type

anime['source'] = anime['source'].replace({'4-koma manga': 'Manga', 'Web manga': 'Manga', 'Digital manga': 'Manga',
                                           'Light novel': 'Novel(Book)', 'Visual novel': 'Novel(Book)',
                                           'Novel': 'Novel(Book)','Picture book': 'Novel(Book)', 'Book': 'Novel(Book)',
                                           'Card game': 'Game', 'Music': 'Other', 'Radio': 'Other'}, regex = True)

# List all resulting distinct source
# Reference: https://www.kaggle.com/xthunder94/category-visualization (same below)
source_list = set()

for entry in anime['source']:
    if not type(entry) is str:
        continue
    source_list.update(entry.split(", "))

print(source_list)
print("Total Sources: " + str(len(source_list)))
# Total Unique Sources: 5
```

```
{'Novel(Book)', 'Manga', 'Game', 'Other', 'Original'}
Total Sources: 5
```

For the 'rating' column, we combine R+ - Mild Nudity, R - 17+ (violence & profanity), Rx - Hentai as one distinct rating as 'R(R+, Rx)'. There are 5 unique rating categories after the reclassification.

```python
# Re-classify / Normalize the 'rating' in anime_sub
# Combine some sub-rating into a parent rating

anime['rating'] = anime['rating'].replace({"R+ - Mild Nudity": 'R',
                                           "R - 17+ (violence & profanity)": 'R',
                                           "Rx - Hentai": 'R'}, regex = False)

# List all resulting distinct rating
rating_list = set()

for entry in anime['rating']:
    if not type(entry) is str:
        continue
    rating_list.update(entry.split(", "))

print(rating_list)
print("Total Ratings: " + str(len(rating_list)))
# Total Unique Ratings: 5
```

```
{'PG-13 - Teens 13 or older', 'R', 'G - All Ages', 'None', 'PG - Children'}
Total Ratings: 5
```

For the 'genre' column, we implement multiple classification on those subgenres.

We combine 'Ecchi' [5], 'Shoujo Ai' [6], 'Yaoi' [7], 'Shounen Ai' [7], 'Yuri' [8], 'Harem' [9], 'Hentai' [10] as one distinct genre 'Hentai'; combine 'Seinen', 'Shounen' as one distinct genre 'Shounen'; combine 'Shoujo', 'Josei' as one distinct genre 'Shoujo'; combine 'Psychological', 'Dementia' as one distinct genre 'Psychological'; combine 'Vampire', 'Demons' as one distinct genre 'Demons'; combine 'Supernatural', 'Super Power', 'Magic' as one distinct genre 'Supernatural'; combine 'Horror', 'Thriller' as one distinct genre 'Horror'. There are 30 unique genre categories after the reclassification.

```python
# Re-classify / Normalize the 'genre' in anime_sub
# Combine some subgenres into a parent genre

genre_replace_dict = {'Ecchi': 'Hentai', 'Shoujo Ai': 'Hentai', 'Yaoi': 'Hentai', 'Shounen Ai': 'Hentai',
                      'Yuri': 'Hentai', 'Harem': 'Hentai', 'Seinen': 'Shounen', 'Josei': 'Shoujo',
                      'Dementia': 'Psychological', 'Vampire': 'Demons', 'Super Power': 'Supernatural',
                      'Magic': 'Supernatural', 'Thriller': 'Horror'}

for key in genre_replace_dict.keys():
    anime['genre'] = anime['genre'].apply(lambda x: x.replace(key, genre_replace_dict[key]))

# Remove the duplicate genre for each anime
for x in anime['genre']:
    anime['genre'] = anime['genre'].apply(lambda x: ', '.join(set(x.split(', '))))

# List all resulting distinct genres
genre_list = set()

for entry in anime['genre']:
    if not type(entry) is str:
        continue
    genre_list.update(entry.split(", "))

print(genre_list)
print("Total Genres: " + str(len(genre_list)))
# Total Unique Genres: 30
```

```
{'Drama', 'Music', 'Martial Arts', 'Supernatural', 'Adventure', 'Romance', 'Space', 'Military', 'School', 'Horror',
'Game', 'Comedy', 'Fantasy', 'Kids', 'Historical', 'Police', 'Cars', 'Parody', 'Slice of Life', 'Demons', 'Psychologi
cal', 'Samurai', 'Action', 'Mecha', 'Sports', 'Shounen', 'Hentai', 'Mystery', 'Shoujo', 'Sci-Fi'}
Total Genres: 30
```

Subtraction

In order to reduce the size of the dataset, we determine unneeded columns to be dropped.

```python
# Determine the unneeded columns to be dropped in each dataframe
anime_dropped_columns = ['title_english', 'title_japanese', 'title_synonyms', 'image_url', 'aired', 'duration',
                         'background', 'premiered', 'broadcast', 'licensor', 'opening_theme', 'ending_theme']

# Drop unneeded columns from dataframe anime
anime_sub = anime.drop(columns = anime_dropped_columns)
```

```python
# Filter out rows where 'premiered_season' is Not Clear
anime_sub = anime_sub[anime_sub['premiered_season'] != 'Not Clear']
```

Data Preprocessing Summary

After completing the data preprocessing procedures, the resulting shape of those three cleaned datasets are:

```python
print('anime:', anime_sub.shape)
print('user:', user.shape)
print('animelists:', animelists.shape)
```

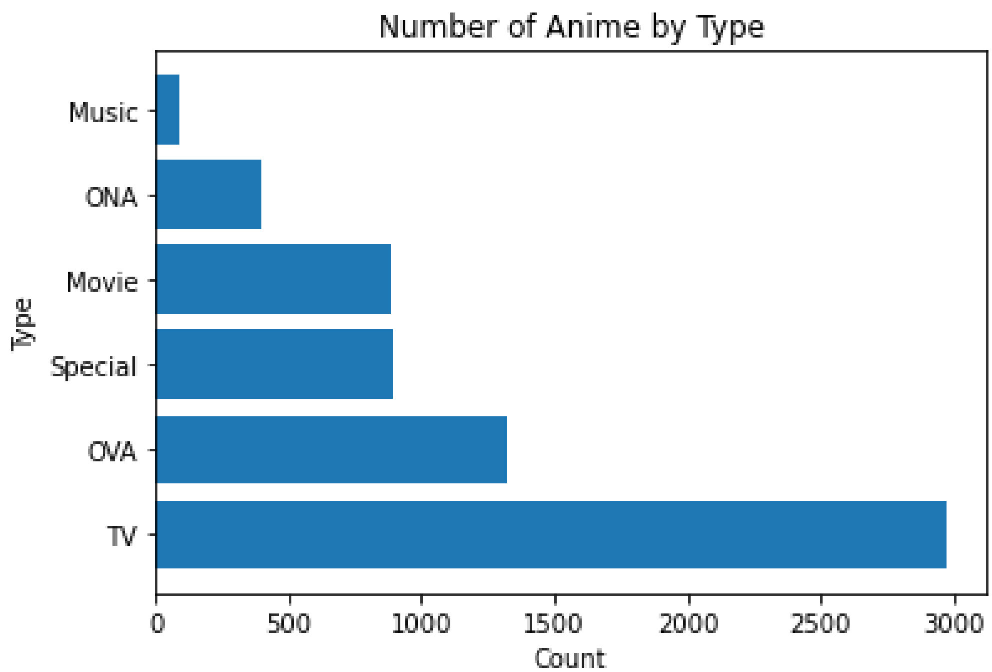
```
anime: (6575, 23)
user: (108705, 17)
animelists: (31283581, 11)
```

Exploratory Data Analysis (EDA)

In order to have a general understanding of the datasets as well as identify some interesting points where we can deep dive into developments questions, we conduct exploratory data analysis.

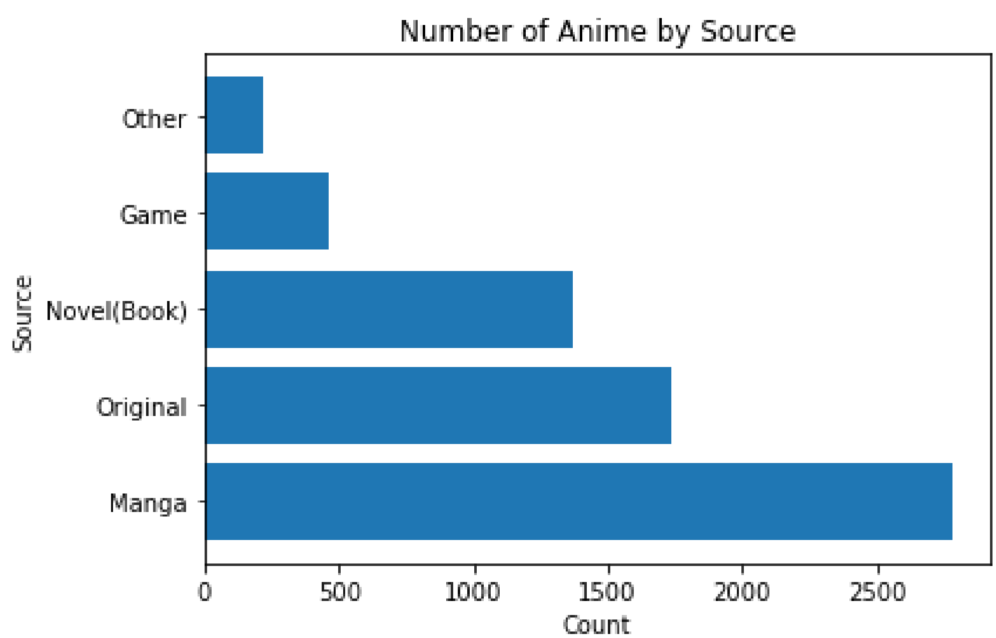
1. **Anime Type**

   There are 5 distinct anime types, which are Music, ONA, Movie, Special, OVA and TV. Among all, TV series is the most common type with nearly 3000 anime works.

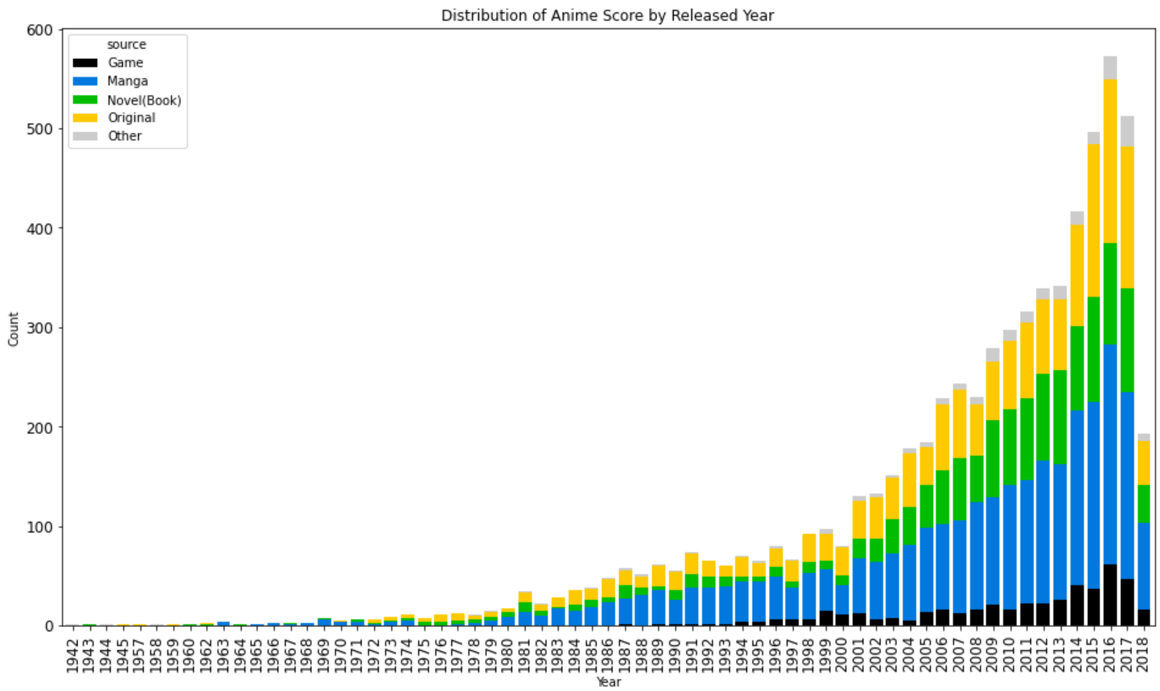


2. **Anime Source**

There are 5 distinct anime sources, which are Manga, Original, Novel(Book), Game and other. Among all, manga is the most common source with over 2500 anime works.

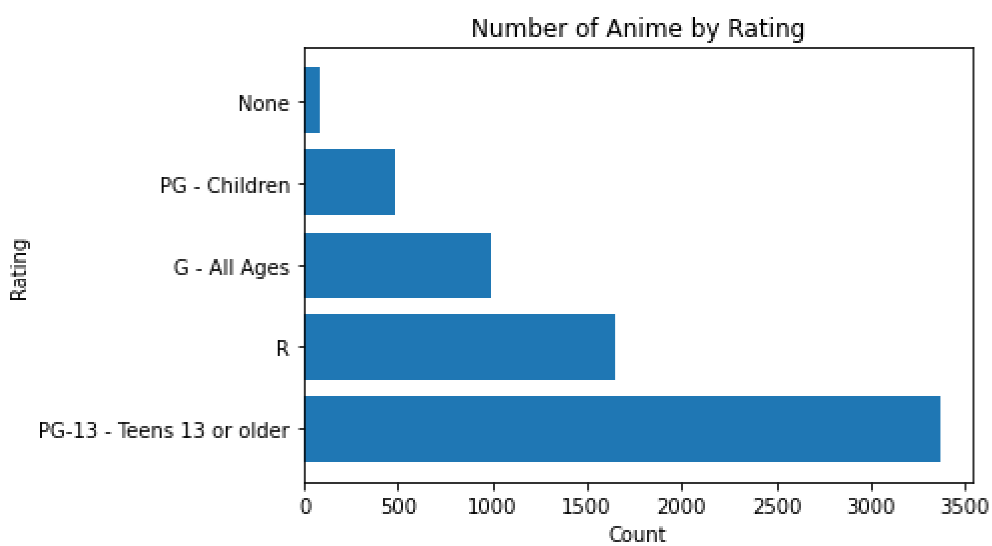


When we plot out the number of each anime source over the anime released year, we surprisingly find out that - although Manga adapted anime maintain the major portion, there are significant increasing trends for original anime and game adapted anime since 2014.



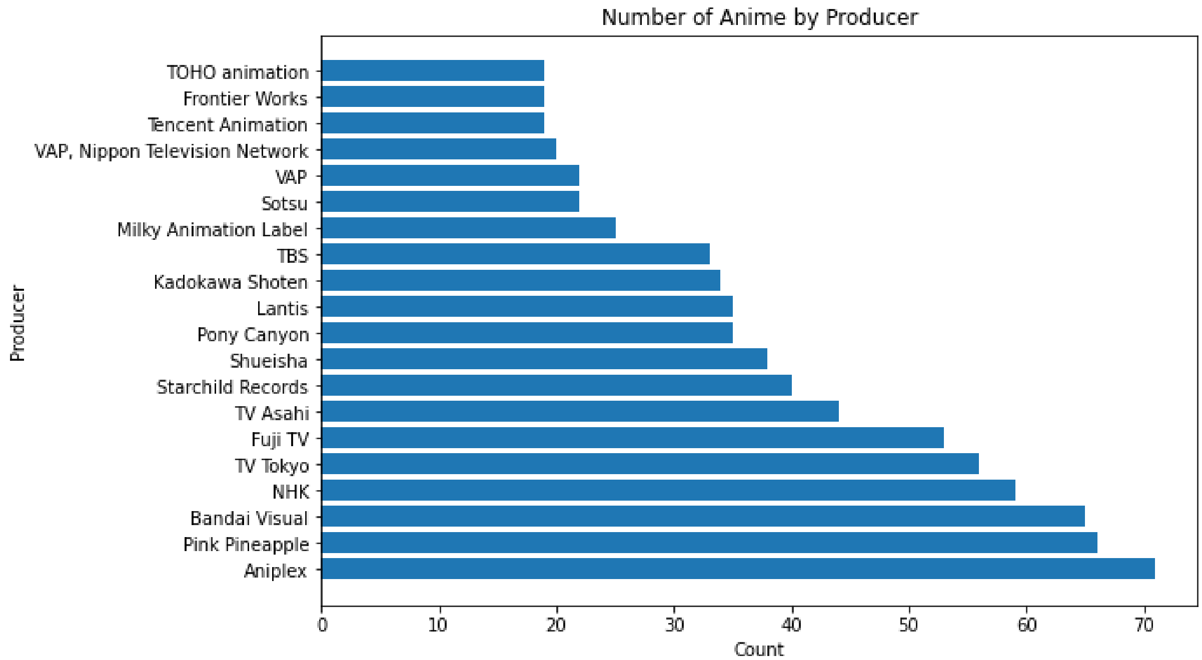
3. **Anime Rating**

   There are 4 distinct anime ratings, which are PG-13 - Teen 13 or older, R, G - All Ages and PG - Children. Among all, PG-13 - Teen 13 or older is the most common rating with nearly 3500 anime works.



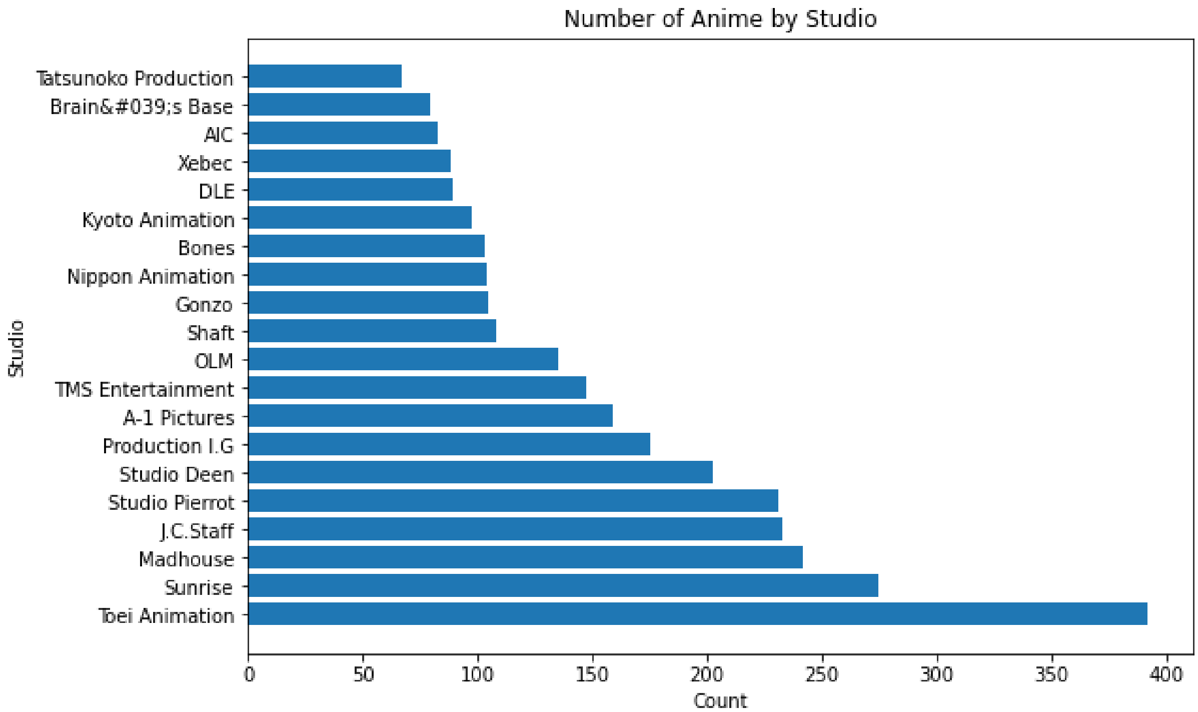
4. **Anime Producer**

   There are 2374 unique producers listed in the 'anime' dataframe. We only sort and display the top 20 producers of the all 2347 by the descending order of the number of anime broadcasted by each producer. According to the barplot below, Aniplex is the one that broadcasts the most anime among all, with over 70 works, followed by Pink Pineapple, Bandal Visual, NHK, TV Tokyo and Fuji TV.



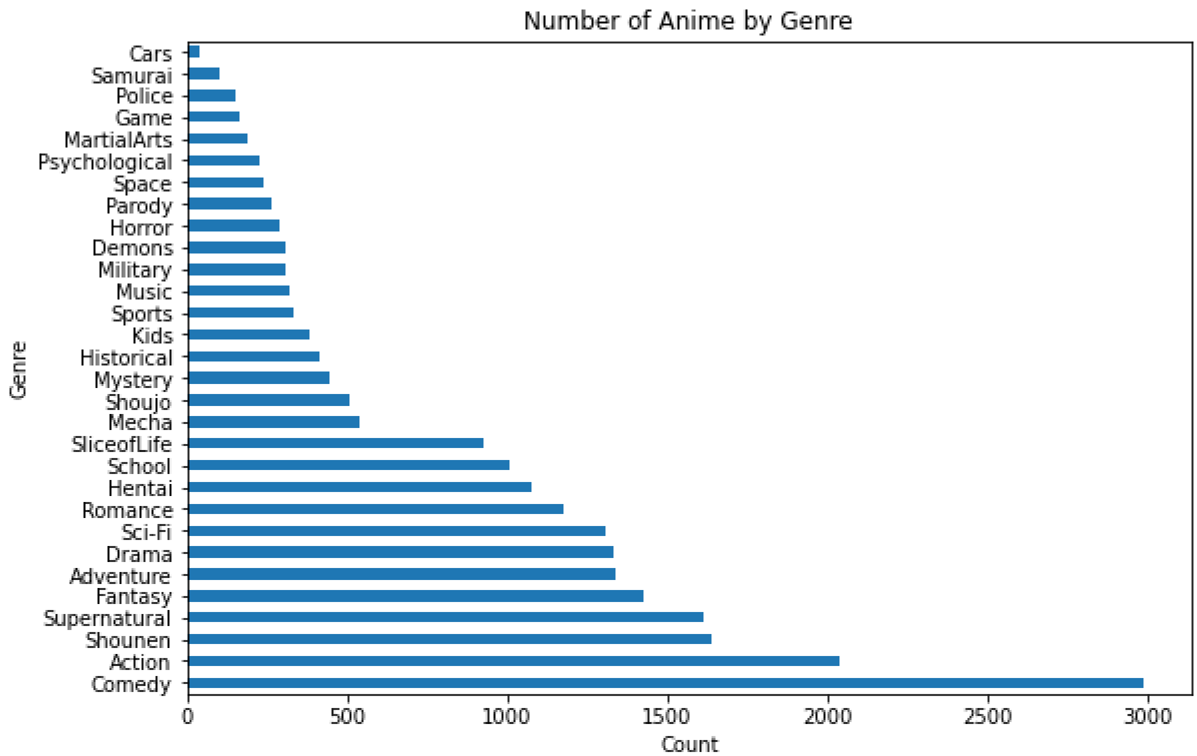
5. **Animation Studio**

There are 707 unique studios listed in the 'anime' dataframe. We only sort and display the top 20 studios of the all 707 by the descending order of the number of anime created by each studio. According to the barplot below, Toei Animation is the one that creates the most anime among all, with over 70 creations, followed by Sunrise, Madhouse, J.C. Staff, Studio Pierrot and Studio Deen. Toei Animation created many great works, including Sailor Moon, Dragon Ball and One Piece [11]



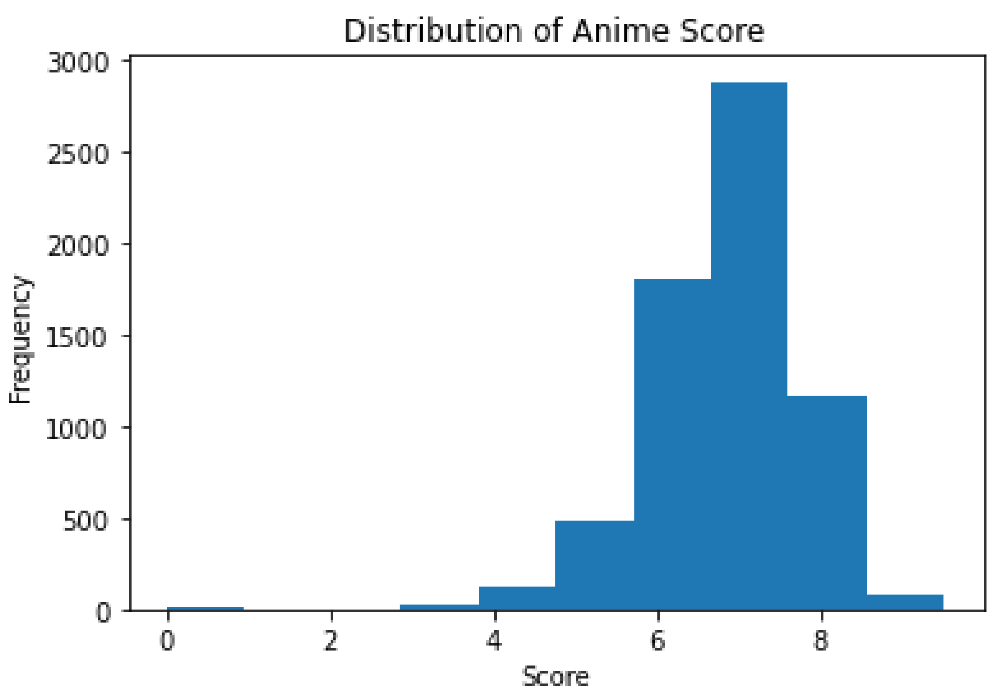
6. Anime Genre

There are 30 unique genres listed in the 'anime' dataframe. We plot out the number of anime of each genre in descending order. According to the barplot below, Comedy is the domain genre - nearly 3000 anime works contain the comedy element. Action, Shounen and Supernatural are also the top genre among all. However, genres like Cars, Samurai and Police are very unpopular, with less than 50 anime works.



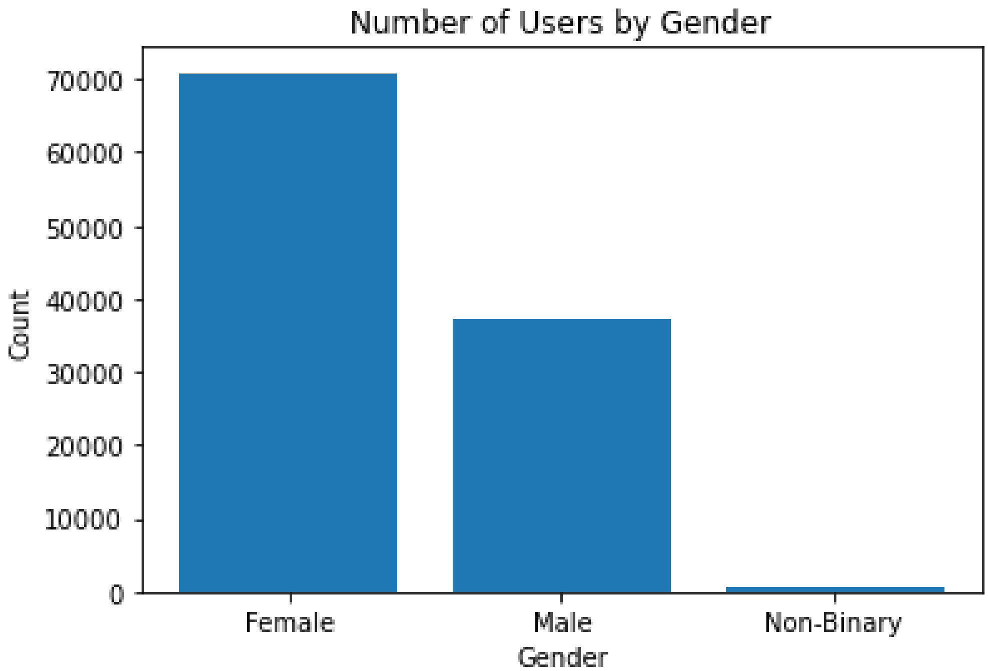
7. **Anime Score**

We plot out the distribution of the anime score. According to the histogram below, overall, most anime scores are within the range of 6 to 8.



8. User Gender

According to the barplot below, overall, there are significantly more female users than males and non-binary users. The number of female users are reaching 70,000, nearly twice as many as male users' scale.



Analytics Questions

With the general understanding of those three datasets, we then design several analytics questions in order to conduct further exploration on the datasets, deep dive into the interesting patterns then generate conclusions based on our findings.

Question 1: Conducted based on time period

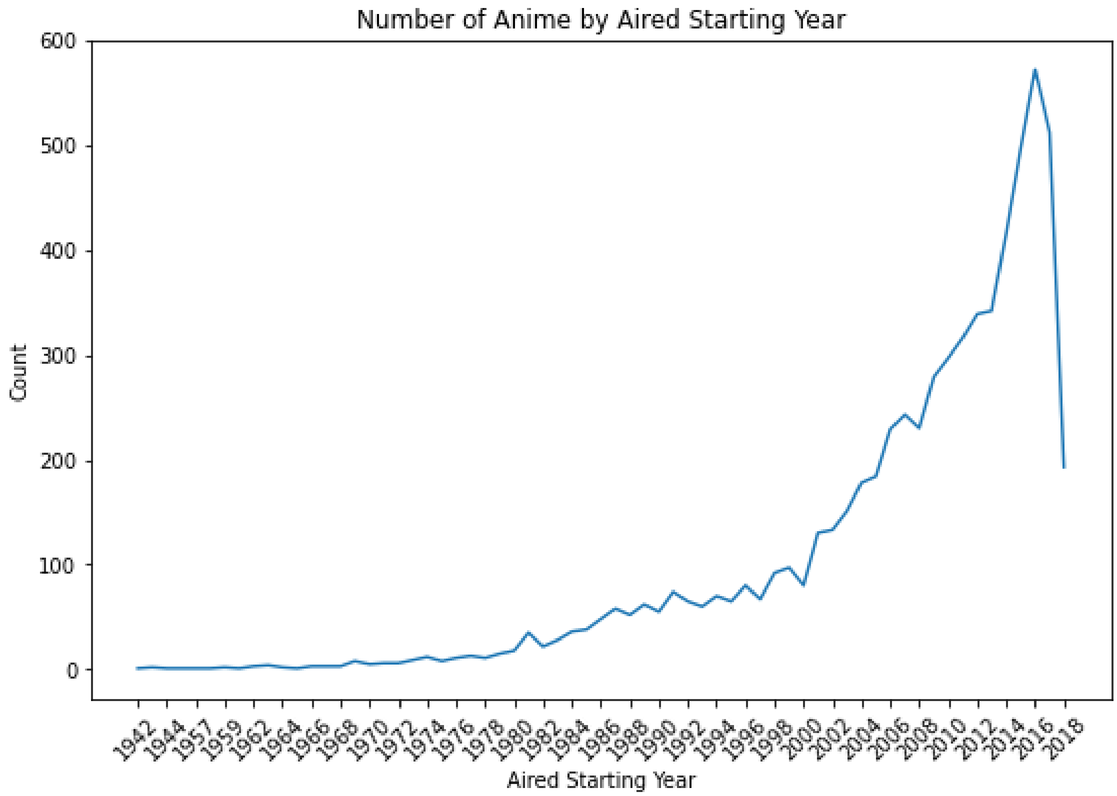
1. Are there any anime features correlated to certain seasons / months? If so, what are those patterns?
2. What are the common combinations of anime genres? Do the combinations vary across the premiered season?

a. Seasonal / Monthly Anime Features

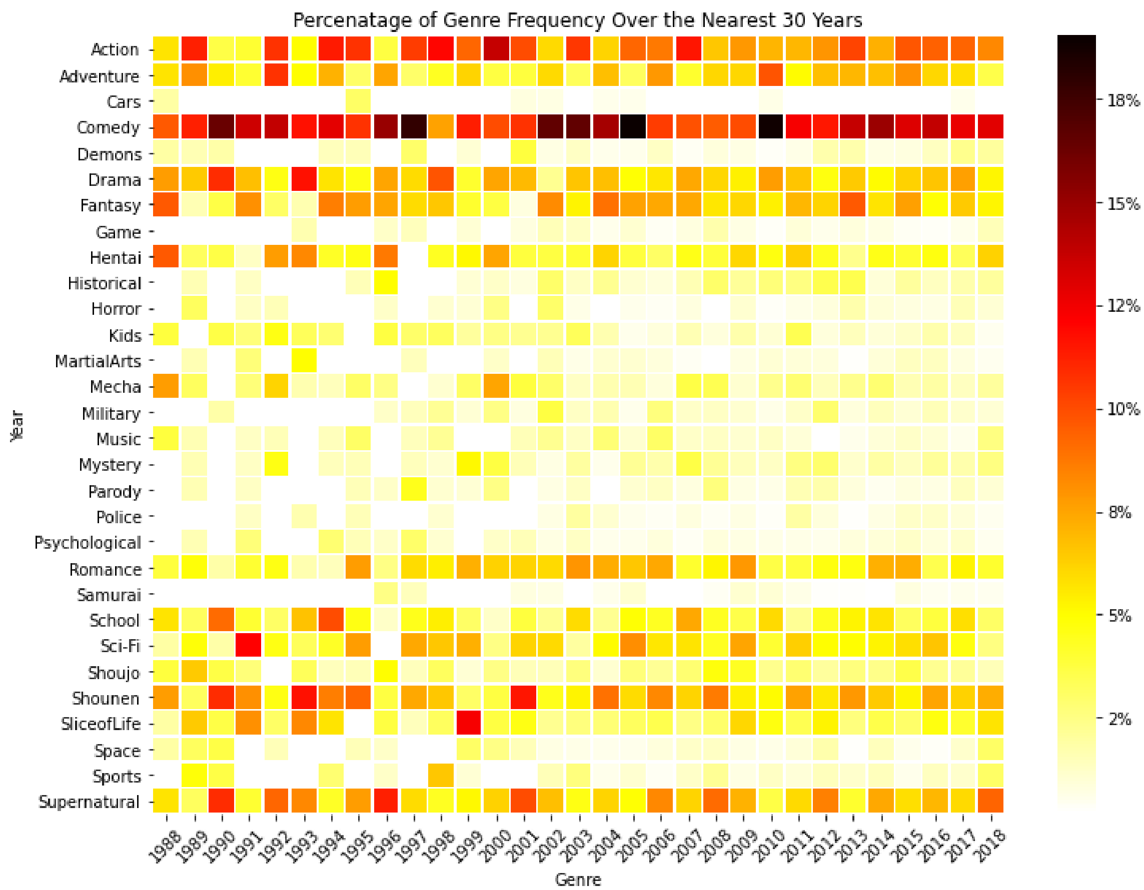
Anime Genre

Firstly, we aggregate the 'anime' dataframe by counting the number of anime by released year, premiered seasons and aired months separately to gain an overview on the trend.

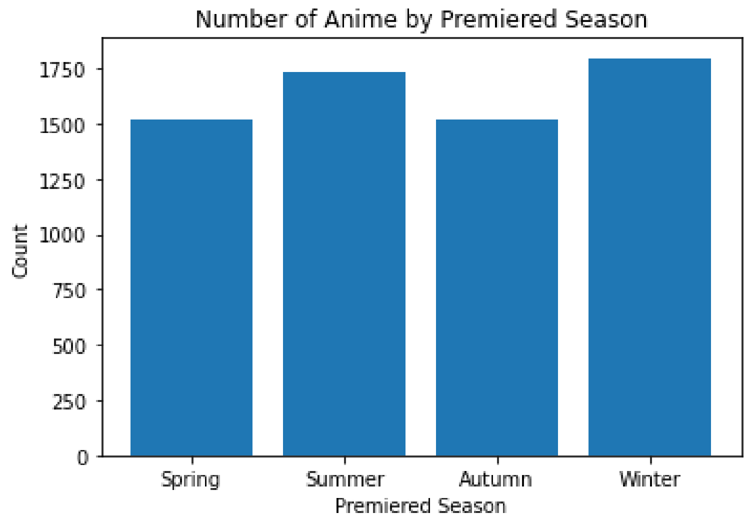
According to the line plot below, it displays the number of anime versus released year. The otaku subculture had expanded its effect around the late 1980s. The 2000s marked a trend of emphasis of the otaku subculture.

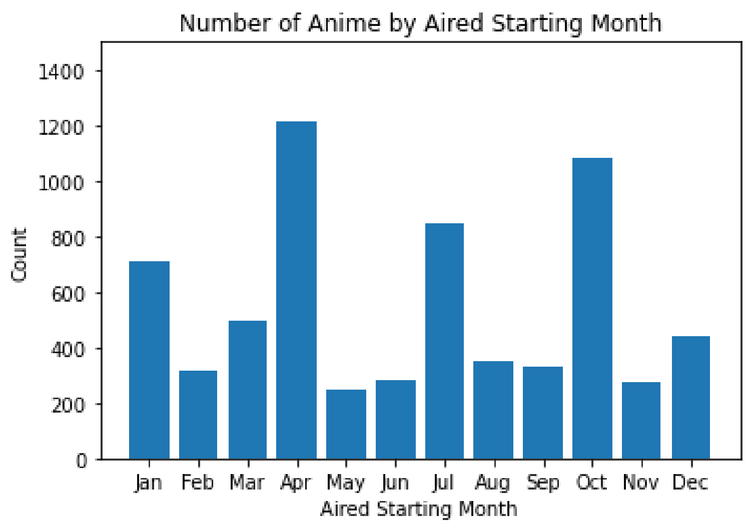


We summarize the anime genre frequency over the last 30 years by calculating the number of anime for each genre by the released year in percentage format. The heat map shows the annual trend of the anime genre from 1988 to 2018. It is not surprising to find out that the Comedy is the hottest genre over the last 30 years, followed by Action, Adventure, Shounen and Supernatural.



These two bar plots below display the number of anime over premiered seasons and aired starting month. Intuitively, there are more anime works released during summer and winter, and to be more specific - more anime works released in April, July and October.





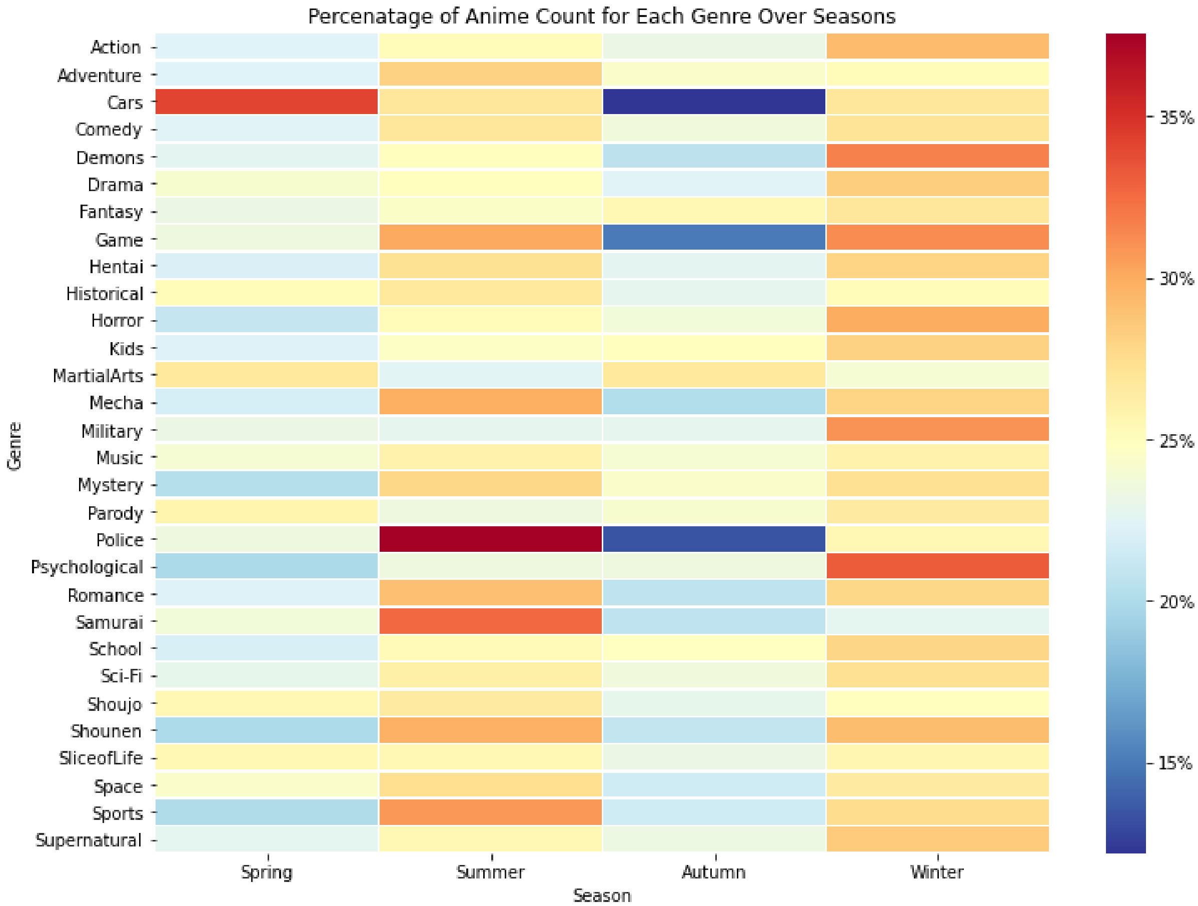
Apart from the plots above, we also generate a summary table that describes the top 10 anime genres overall and the seasonal trend, count in actual number. According to the summary table below, the hottest genres like comedy and action do not have specific seasonal variance based on the actual number of animes. Their popularities are always! In general, the top 10 anime genre categories are similar over season but the relative popularities are in different order. For example, the School genre is one of the top 10 common genres in autumn, which is relatively unpopular in the other three seasons.

```
The top 10 Anime Genre and Count by Season:
```

	Spring	Summer	Autumn	Winter	Overall
0	(Comedy, 669)	(Comedy, 800)	(Comedy, 706)	(Comedy, 810)	(Comedy, 2985)
1	(Action, 455)	(Action, 511)	(Action, 472)	(Action, 596)	(Action, 2034)
2	(Supernatural, 366)	(Shounen, 488)	(Supernatural, 375)	(Shounen, 478)	(Shounen, 1636)
3	(Fantasy, 331)	(Supernatural, 410)	(Fantasy, 362)	(Supernatural, 458)	(Supernatural, 1609)
4	(Shounen, 327)	(Adventure, 375)	(Shounen, 343)	(Fantasy, 383)	(Fantasy, 1426)
5	(Drama, 323)	(Fantasy, 350)	(Adventure, 326)	(Drama, 376)	(Adventure, 1335)
6	(Sci-Fi, 300)	(Romance, 342)	(Sci-Fi, 309)	(Sci-Fi, 357)	(Drama, 1330)
7	(Adventure, 299)	(Sci-Fi, 342)	(Drama, 299)	(Adventure, 335)	(Sci-Fi, 1308)
8	(Romance, 262)	(Drama, 332)	(School, 249)	(Romance, 327)	(Romance, 1175)
9	(Hentai, 238)	(Hentai, 293)	(Romance, 244)	(Hentai, 302)	(Hentai, 1077)

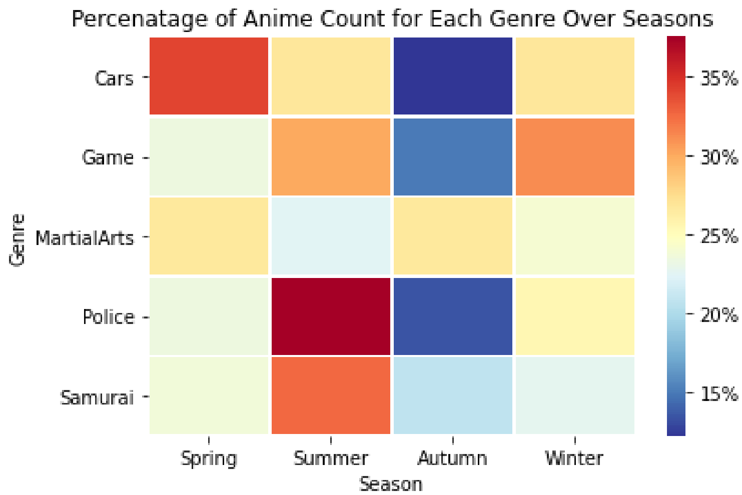
Then, we visualize the season trend of each anime genre using a heatmap. Each cell represents the number of anime per genre per season divided by the number of anime count within the genre subtotal in percentage format. More red represents more anime work of that genre that was released in that season; more blue represents less.

According to the heat map below, most anime regardless of genre are mostly released during both summer and winter. However, As for the 5 least popular genres - cars, game, martial arts, police, samurai - they have significant seasonal variance. Anime of those genres tend to only focus on one releasing season - summer or winter; or rather target at spring and autumn to avoid the anime releasing crowds.



Based on the overall seasonal trend for each genre displayed above, we subtract those 5 least popular genres out of the total, showing their seasonal number of anime works counts together with the average score for each genre per season in order to analyze how the anime audiences respond. Overall, these unpopular genres perform fairly well, with average scores around 7.

Although anime of cars genre mostly targets at spring release, the response is not better than those released in summer. Although the police genre and samurai genre mainly focus on summer release, there are potential positive responses during autumn and winter respectively.




```
The average score of the 5 least popular genre by season:
```

	Spring	Summer	Autumn	Winter
Cars	6.743571	7.459091	6.772000	6.883636
Game	6.927949	6.784000	6.892400	6.849038
Martial Arts	6.875882	6.999535	6.881569	6.986304
Police	6.855143	7.433750	7.198000	7.025263
Samurai	7.270000	7.235455	7.199524	7.343478

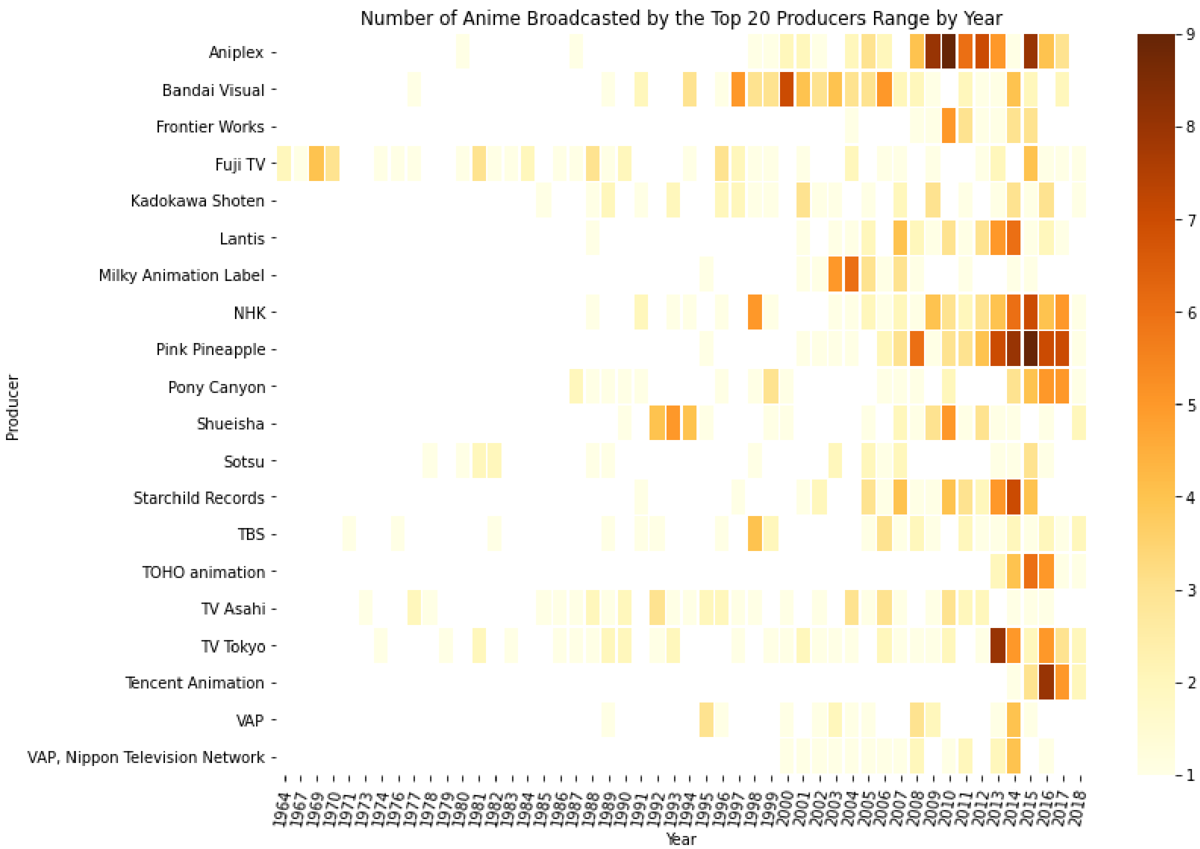
b. Anime Producer

As we summarize the number of anime broadcasted by the top 20 producers over the released year, we could find out some interesting patterns.

Fuji TV is one of the oldest producers to broadcast anime since 1964. Before 1997, the animation market was starting up, each producer produced less than 5 anime on average. During this time, Fuji TV (before 1990s) and Shueisha (1992-1995) were the two main producers.

During 1997-2007, the animation market witnessed the first boost in broadcasting anime works in some producers, such as Bandai Visual (1997, 2000, 2007), Milky Animation Label (2003-2004), NHK (2005).

After 2007: there was a second boost when most of the top 20 producers began to boost the broadcasting scale. Aniplex(2009-2016, most in 2010); lantis(2007-2014); NHK(since 2009 increasing trend); Pink Pineapple(since 2008 increasing trend, most in 2015); Pony Canyon(2014-2018); Shueisha(few market share); Starchild Records(since 2008 increasing trend, most in 2014); TOHO Animation(2014-2016); TV tokyo(2013-2016, most in 2013); Tencent Animation(2015-2017, most in 2016).



c. **Animation Studio**

We summarize the annual trend for the number of anime creations produced by each one of the top 20 studios using heatmap. According to the heatmap below:

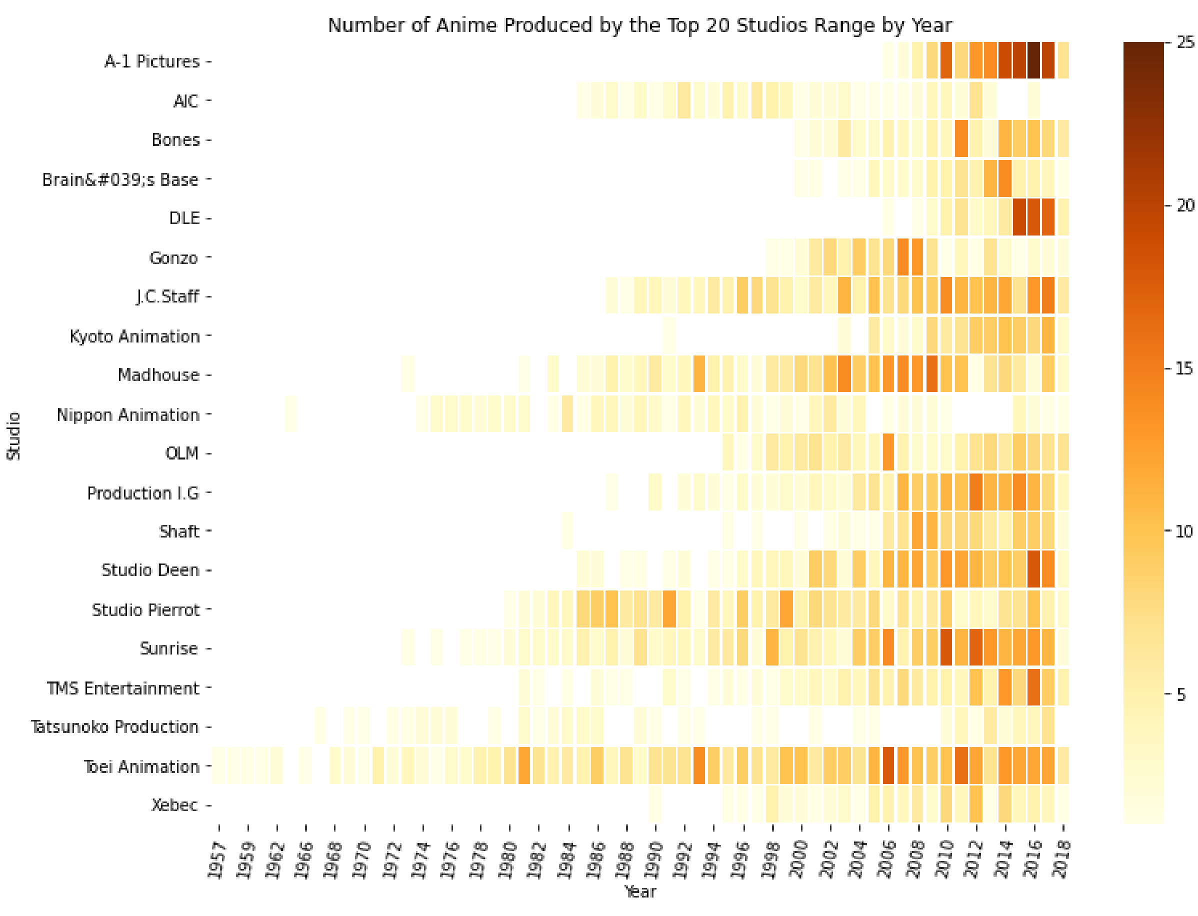
The first tier's studios produced over 150 anime works.

1. Sunrise: rapid developed since 2004
2. Madhouse: rapid developed 1998-2013, its market share shrinked a little bit after 2012
3. J.C.Staff: maintained steady increasing development since 1988
4. Studio Pierrot: maintained a good portion of market share overall
5. Studio Deen: maintained rapid development since 2000 till now
6. Production I.G: maintained steady increasing development since 1988

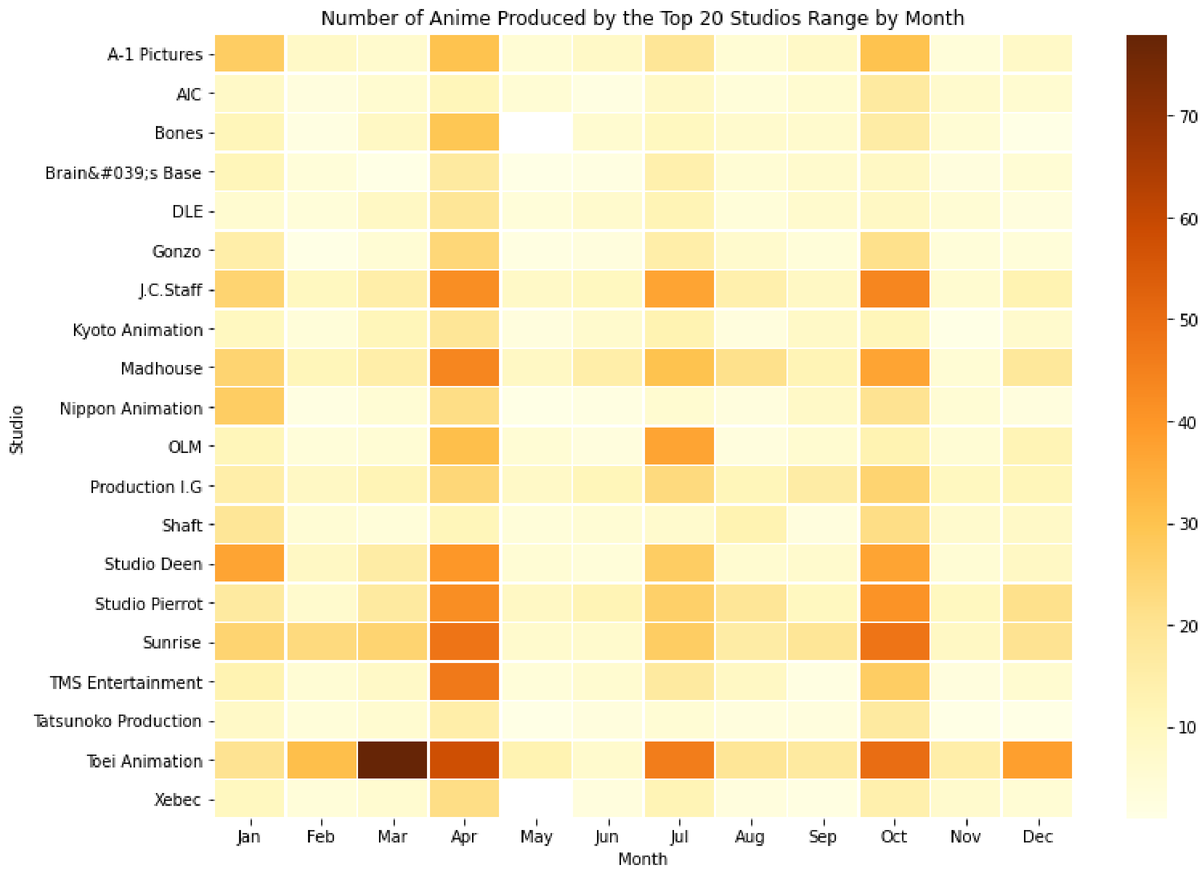
7. A-1 Pictures: maintained rapid development since founded in 2005 till now, ate up the most market share during 2014-2017

The second tier's studios produced 75-150 anime works.

1. TMS Entertainment: maintained steady increasing development since 1988
2. OLM: considerable market share in 2006, after then its market share shrinked
3. Sharft: developed since 2006
4. Gonzo: 2000-2009, after then the portion shrink
5. Nippon Animation: maintain a fixed production scale around 3-4 anime per year, shrinked after 2000
6. Bones: maintained steady increasing development since founded in 1998, good market share in 2011
7. Kyoto Animation: maintained a fixed production scale around 10 anime per year since 2008
8. DLE: achieve high production during 2015-2017



Toei Animation is the biggest and one of the oldest studios, producing 392 anime(1981, 1993, 2006, 2011, steady increasing trend in market share); most in spring(March), then winter(October, December), fewer in summer(April) and autumn(July). In general, studios tend to release the anime in April, July, October. However, Toei Animation mainly targets at March thus eats up the market share in spring, same for December.



d. Seasonal / Monthly Mix and Match of Anime Genres

In the 'anime' dataframe, for each anime, there are multiple genre tags in a single 'genre' column. Therefore, we generate a bigram using nltk package to explore the most common anime genre combination over seasons. More frequent of the genre pairs, higher the bigram score, thus giving out a higher rank.

According to the resulting bigram objects, comedy is the panacea, since it is the most popular genre in general. Half of the overall top 10 anime genre pairs contains a Comedy genre. All four seasons share the same most common genre combination: Action and Science-Fiction.

Another interesting fact is that the combination of Slice of life genre and Comedy genre is relatively more popular during spring and summer than that in autumn and winter when the combination of Adventure and Action is more common.

Compared with the overall common genre combinations, there are also a couple of popular genre combinations for specific seasons, such as more combination of Science-Fiction & Comedy in spring; more combination of Shounen and Adventure in summer; more combination of School and Comedy in autumn; less combination of Comedy and Action in summer and autumn.

```
The top 10 Anime Genre Pairs by Season:
```

	Spring	Summer	Autumn	Winter	Overall
0	(action, sci-fi)	(action, sci-fi)	(action, sci-fi)	(action, sci-fi)	(action, sci-fi)
1	(life, comedy)	(life, comedy)	(adventure, action)	(adventure, action)	(adventure, action)
2	(adventure, action)	(adventure, action)	(life, comedy)	(supernatural, fantasy)	(life, comedy)
3	(supernatural, fantasy)	(supernatural, fantasy)	(supernatural, fantasy)	(life, comedy)	(supernatural, fantasy)
4	(comedy, action)	(comedy, hentai)	(comedy, hentai)	(comedy, action)	(comedy, action)
5	(comedy, supernatural)	(comedy, supernatural)	(school, comedy)	(comedy, supernatural)	(comedy, hentai)
6	(romance, school)	(comedy, action)	(romance, school)	(romance, school)	(romance, school)
7	(sci-fi, comedy)	(romance, school)	(comedy, action)	(comedy, hentai)	(comedy, supernatural)
8	(comedy, hentai)	(fantasy, shounen)	(comedy, supernatural)	(fantasy, shounen)	(fantasy, shounen)
9	(fantasy, shounen)	(shounen, adventure)	(action, comedy)	(action, comedy)	(school, comedy)

Question 2: Conducted based on users' characteristics.

How do users' watching and rating preferences vary by gender?

1. watching behavior vary by gender

   To analyze users' watching behavior in general, We grouped 'user_days_spent_watching' by gender and counted the mean values. We found females spent less average days watching anime than Male did. Females' average watching days is 93, while Males' average watching days is approximately 121 days.

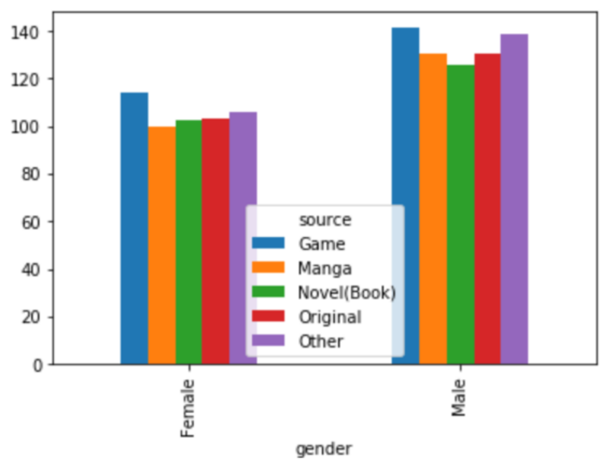
```
# the average watching days of different gender

watching_gender=merged_list.groupby('gender')['user_days_spent_watching'].mean()
watching_gender
```

```
gender
Female     93.007210
Male      121.710607
Name: user_days_spent_watching, dtype: float64
```

Next, we also grouped 'user_days_spent_watching' by gender as well as by source of anime and counted mean values. There are six kinds of source of anime in our data, which are Game, Manga, Novel, Original and Other. From the pivot table below, we can see that for both female and male, the highest bar is Game. Therefore, Game adaptation anime may have a great potential market. People are interested in these types of anime because they tend to spend more days to enjoy watching.

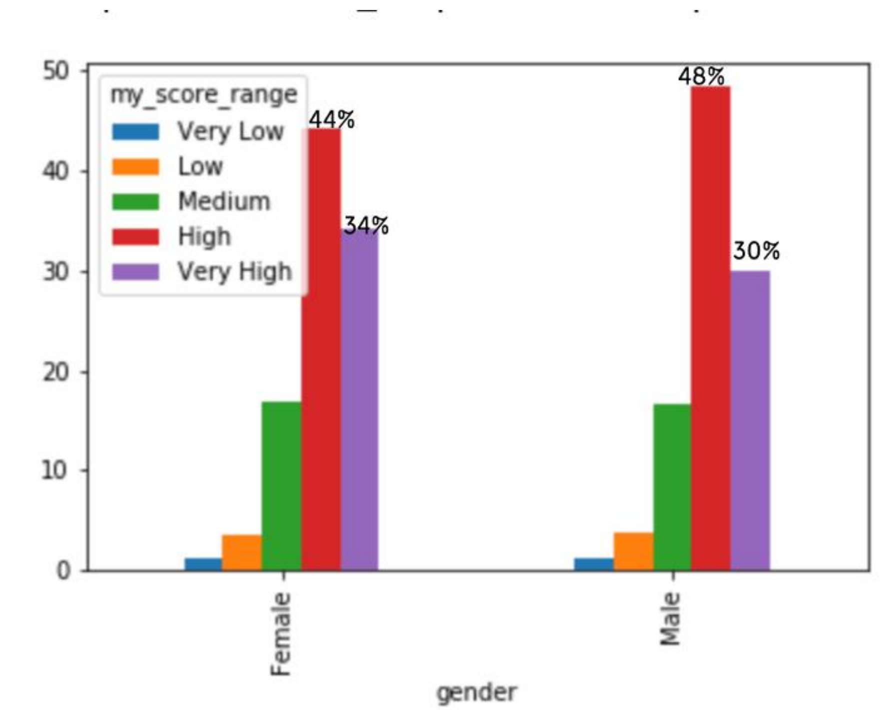
```
: watching_gender_source.pivot('gender','source','mean').plot(kind='bar')
```

```
: <matplotlib.axes._subplots.AxesSubplot at 0x7ff7cf3b32d0>
```



2. Score behavior vary by gender

   To analyze score behavior, we set several score levels based on the 'my_score' column in data. 'my_score' is a column of rating score from users. Those scores are in the range of 0-10. We assumed that 0 score represents null rating or extreme value, so we filtered 0 score and set a range from 1-10. We divided the score 9 and 10 into 'Very High' level, 7 and 8 as 'High', 5 and 6 as 'Medium', 3 and 4 as 'Low', 1 and 2 as 'Very Low'. There are five levels in total. After that, we calculated the frequency of each level for female and for male respectively. The frequency bar chart is shown below. We can conclude that both female and male prefer to rate 7-8 high level and female is more possible to rate in very high level.



We also found the top 10 scored animes based on 'my_score' in descending order. For those top 10 animes, we listed their genres grouped by gender as below:

```
F_top10['genre'].tolist()
```

```
['Comedy',
 'Supernatural, Parody',
 'Sports, Shounen',
 'Comedy',
 'Kids, Adventure, Comedy',
 'Kids, Adventure, Action',
 'Supernatural, Fantasy, Slice of Life',
 'Mecha, Kids, School, Action',
 'Sci-Fi',
 'Samurai, Parody, Action, Sci-Fi, Comedy, Historical, Shounen']
```

Female top 10 anime genres

```
M_top10 = pd.merge(M_scored,anime1[['title','genre']], how='inner')
M_top10['genre'].tolist()
```

```
['Action, Drama, Fantasy, Adventure, Comedy, Supernatural, Military,
 'Horror, Sci-Fi',
 'Slice of Life, Romance, Drama, Comedy, Supernatural',
 'Samurai, Parody, Action, Sci-Fi, Comedy, Historical, Shounen',
 'Supernatural, Drama, School, Romance',
 'Samurai, Parody, Action, Sci-Fi, Comedy, Historical, Shounen',
 'Supernatural, Action, Adventure, Shounen',
 'Samurai, Parody, Action, Sci-Fi, Comedy, Historical, Shounen',
 'Horror, Sci-Fi',
 'Drama, School, Shounen',
 'Samurai, Parody, Action, Sci-Fi, Comedy, Historical, Shounen',
 'Supernatural, Demons, Mystery, Comedy',
 'Drama, Slice of Life, Game, Shounen']
```

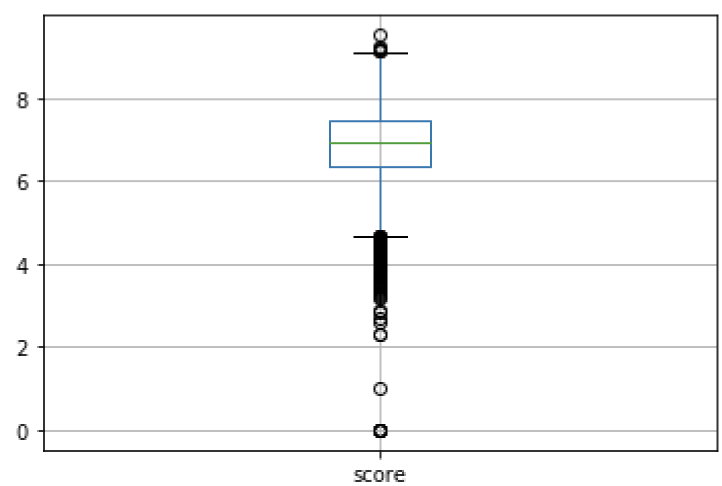
Male top 10 anime genres

From those genres, we can find different anime type preferences of two genders. Females prefer Comedy, Kids and Action, while male prefer Comedy, Action, Sci-Fi, Drama, Supernatural, Historical and Shounen. Comedy and Action are everyone's favorite.

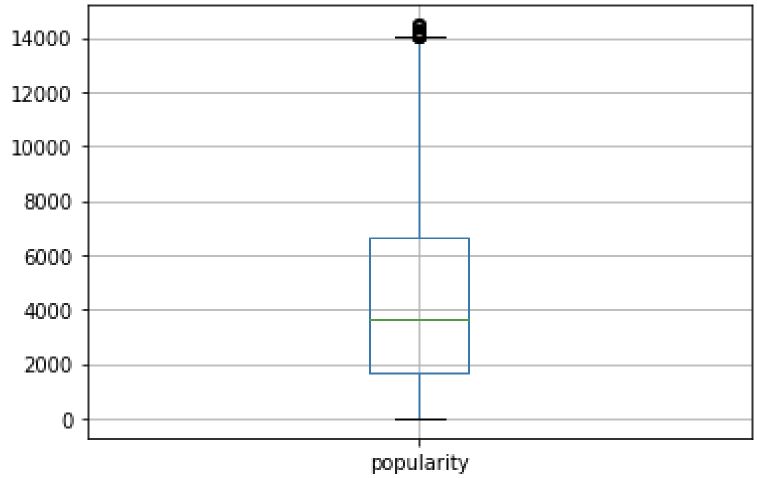
Question 3: Based on the cross analysis on anime scores and popularity

1. What's the difference between animes with both high score and high popularity and animes with high score but relatively low popularity?

    1. **To compare different groups of animes, we need to figure out the threshold to split the whole dataset. For this question, we only need to look at two columns: score and popularity. We used boxplot to show the distribution of these two columns.**



we had extremely low scores below 5 points and extremely high scores above 8.5. For those animes above 8.5 are our research objects. We want to see why those anime were so popular? What genres do those anime belong to? and so on.



For the 'popularity' column, the lower the popularity of anime, the more people love to see this anime. Therefore we don't want to research the animes out of 14000, we focused on those anime popularity below 2000.

2. Animes with high score and high popularity

anime_id	title	episodes	airing	score	scored_by	rank	popularity	members	favorites	duration_min
1535	Death Note	37	False	8.67	1009477	51.0	1	1456378	88696	23.0
5114	Fullmetal Alchemist: Brotherhood	64	False	9.25	733592	1.0	4	1199091	106895	24.0
30276	One Punch Man	12	False	8.73	691845	44.0	5	1020754	30747	24.0
9253	Steins;Gate	24	False	9.14	563857	5.0	8	1010330	92423	24.0
1575	Code Geass: Hangyaku no Lelouch	25	False	8.79	627740	30.0	9	986897	63614	24.0
2904	Code Geass: Hangyaku no Lelouch R2	25	False	8.95	543904	18.0	22	791396	44230	24.0
2001	Tengen Toppa Gurren Lagann	27	False	8.74	449656	41.0	24	787535	50040	24.0
32281	Kimi no Na wa.	1	False	9.19	471398	2.0	33	730076	34912	106.0
11061	Hunter x Hunter (2011)	148	False	9.11	403377	8.0	35	720920	64375	23.0
21	One Piece	0	True	8.54	423868	91.0	35	720133	69760	24.0

	title	episodes	airing	score	scored_by	rank	popularity	members	favorites	duration_min
3800	Death Note	37	False	8.67	1009477	51.0	1	1456378	88696	23.0
1261	Fullmetal Alchemist: Brotherhood	64	False	9.25	733592	1.0	4	1199091	106895	24.0
4213	One Punch Man	12	False	8.73	691845	44.0	5	1020754	30747	24.0
1475	Steins;Gate	24	False	9.14	563857	5.0	8	1010330	92423	24.0
6577	Code Geass: Hangyaku no Lelouch	25	False	8.79	627740	30.0	9	986897	63614	24.0
3548	Code Geass: Hangyaku no Lelouch R2	25	False	8.95	543904	18.0	22	791396	44230	24.0
5235	Tengen Toppa Gurren Lagann	27	False	8.74	449656	41.0	24	787535	50040	24.0
529	Kimi no Na wa.	1	False	9.19	471398	2.0	33	730076	34912	106.0
34	One Piece	0	True	8.54	423868	91.0	35	720133	69760	24.0
223	Hunter x Hunter (2011)	148	False	9.11	403377	8.0	35	720920	64375	23.0

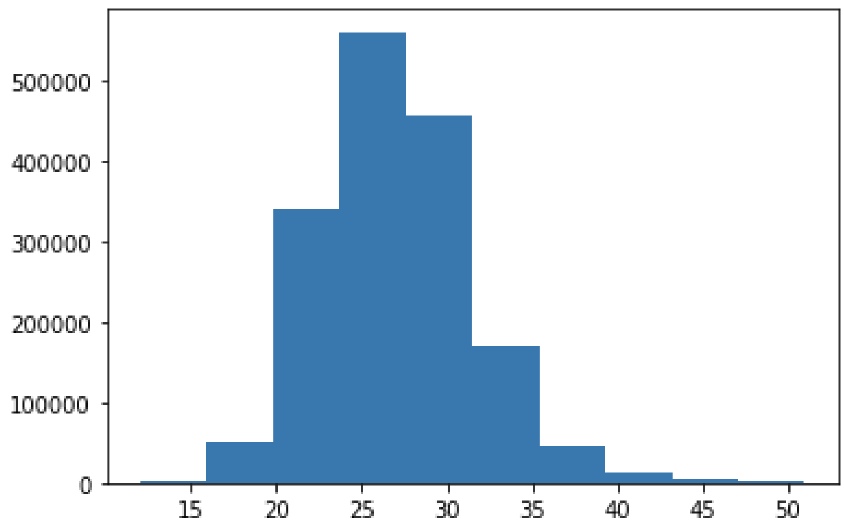
- In the top ten list, except One Piece is still playing, the rest have finished all episodes.
- in the top ten list, except Kimi no Na wa is movie, the rest of the list are TV series

3. Correlation Matrix:

	episodes	airing	score	scored_by	rank	popularity	members	favorites	duration_min
episodes	1.000000	-0.066069	0.217416	0.042950	-0.156640	-0.131980	0.115682	0.255011	-0.366206
airing	-0.066069	1.000000	0.013120	-0.083388	-0.010695	0.000075	-0.025913	0.024217	-0.111503
score	0.217416	0.013120	1.000000	0.219979	-0.946148	-0.215701	0.251364	0.357100	0.027719
scored_by	0.042950	-0.083388	0.219979	1.000000	-0.210494	-0.657299	0.983097	0.847955	0.007338
rank	-0.156640	-0.010695	-0.946148	-0.210494	1.000000	0.234182	-0.234944	-0.298800	-0.034210
popularity	-0.131980	0.000075	-0.215701	-0.657299	0.234182	1.000000	-0.718152	-0.512254	-0.048513
members	0.115682	-0.025913	0.251364	0.983097	-0.234944	-0.718152	1.000000	0.880279	-0.039797
favorites	0.255011	0.024217	0.357100	0.847955	-0.298800	-0.512254	0.880279	1.000000	-0.136432
duration_min	-0.366206	-0.111503	0.027719	0.007338	-0.034210	-0.048513	-0.039797	-0.136432	1.000000

High correlation among score and rank, scored_by and members, scored_by and favorites, popularity and members, members and favorite. Therefore we can say that higher the score, the more people watched this anime, causing higher rank and higher popularity.

4. Distribution of age of people who loved to watch high score and high popularity anime. Mostly young people, the average is about 27 years old. Because the age distribution is normal among all kinds of groups of anime, not much more to repeat below



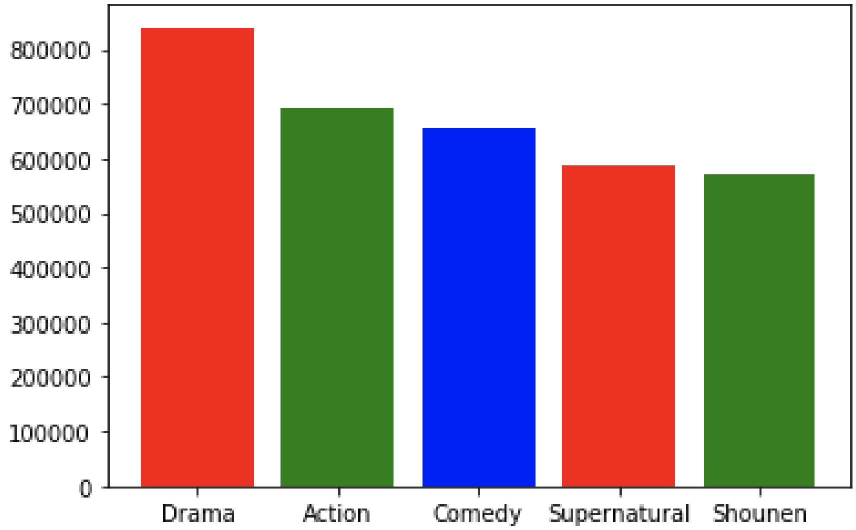
5. Top 10 'my_watched' episodes

anime_id	title	my_watched_episodes	my_rewatching	episodes	score	scored_by	rank	popularity	favorites
1535	Death Note	36.12	0.00	37.0	8.67	1009477.0	51.0	1.0	88696.0
5114	Fullmetal Alchemist: Brotherhood	58.94	0.01	64.0	9.25	733592.0	1.0	4.0	106895.0
30276	One Punch Man	11.60	0.00	12.0	8.73	691845.0	44.0	5.0	30747.0
9253	Steins;Gate	22.59	0.00	24.0	9.14	563857.0	5.0	8.0	92423.0
1575	Code Geass: Hangyaku no Lelouch	24.26	0.00	25.0	8.79	627740.0	30.0	9.0	63614.0
2904	Code Geass: Hangyaku no Lelouch R2	24.53	0.00	25.0	8.95	543904.0	18.0	22.0	44230.0
2001	Tengen Toppa Gurren Lagann	25.56	0.00	27.0	8.74	449656.0	41.0	24.0	50040.0
32281	Kimi no Na wa.	0.98	0.00	1.0	9.19	471398.0	2.0	33.0	34912.0
11061	Hunter x Hunter (2011)	129.54	0.01	148.0	9.11	403377.0	8.0	35.0	64375.0
21	One Piece	498.47	0.00	0.0	8.54	423868.0	91.0	35.0	69760.0

For most high score and high popularity animes, people loved to watch all the episodes one time in general, not so much people love to rewatch them.

6. Quality research: Genre

Drama, Action, Comedy, Supernatural and Shounen are the most popular



7. **Quality research: Studio**

```
studio
Madhouse                    12
Sunrise                      9
Artland                      6
Production I.G               6
A-1 Pictures                 5
Shaft                        5
Bandai Namco Pictures        5
Kyoto Animation              5
```

Madhouse and Sunrise are the main force to product high quality and high score animes

8. **Quality research: Type, high score and high popularity animes are mostly TV series**

```
type
TV         1242016
Movie       348143
OVA          44296
Special       9815
```

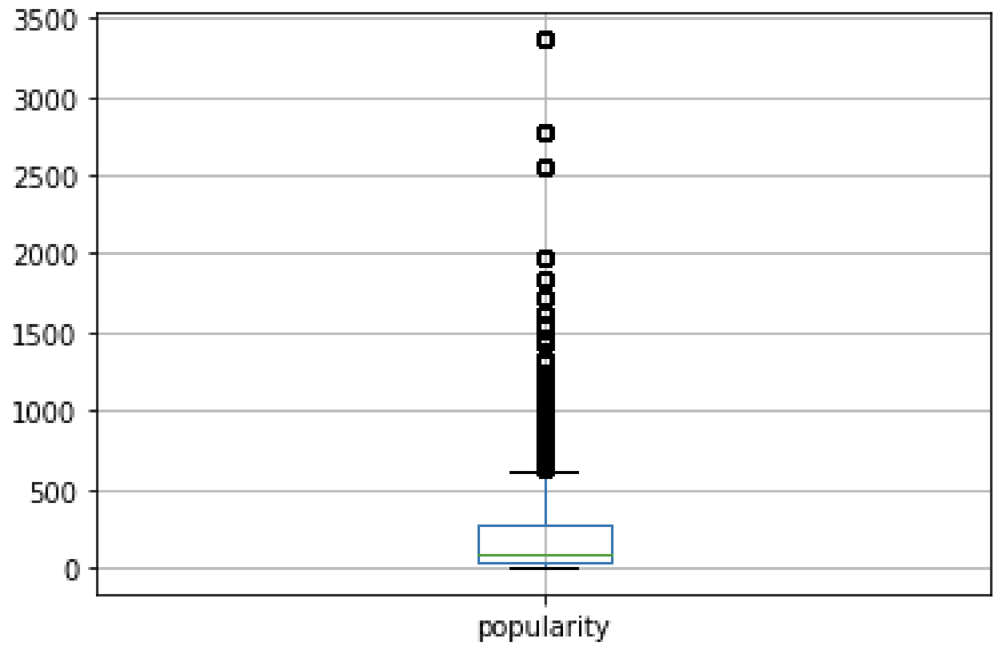
9. **Quality research: Year. 2006 is the year of high score animation production**

aired_from_year	
2006	183755
2011	180210
2016	155910
2015	124945
2014	107598
2008	103120
2009	99758
2004	92966
2017	84825
2012	81096
1999	62845
2001	49879
2013	49393
2010	46981
2018	46471
2007	40789
1997	35809
1998	34651
1988	24710
2005	18937
2000	12462
1993	7160

title	
Death Note	70764
Code Geass: Hangyaku no Lelouch	55829
Hellsing Ultimate	20651
Gintama	20354
Nana	16157

2. Why do high scores but low popularity situations exist?

   For high score anime, the distribution of popularity is like this. Popularity above 600 is relatively low popularity.



First 10 animes that have high score but low popularity

```
title
Kara no Kyoukai 7: Satsujin Kousatsu (Kou)                  11378
Natsume Yuujinchou Shi                                       9047
Hajime no Ippo: New Challenger                               9039
Gintama&#039;: Enchousen                                     8178
Tengen Toppa Gurren Lagann Movie 2: Lagann-hen               7920
Gintama Movie 2: Kanketsu-hen - Yorozuya yo Eien Nare        7521
Slam Dunk                                                    7160
Gintama Movie 1: Shinyaku Benizakura-hen                     6678
Kizumonogatari II: Nekketsu-hen                              6601
Hajime no Ippo: Rising                                       6096
```

a. **The first reason is that most animes are movie animes. Not like TV series, they can't keep hot**

Gintama Movie 1: Shinyaku Benizakura-hen	Manga	Movie	1.0	8.57	42409.0	1138.0	324.0	27.039832
Gintama Movie 2: Kanketsu-hen - Yorozuya yo Eien Nare	Manga	Movie	1.0	9.04	61010.0	826.0	1395.0	26.421487
Gintama': Enchousen	Manga	TV	13.0	9.07	63468.0	685.0	1509.0	26.656884
Gintama.	Manga	TV	12.0	9.02	40441.0	934.0	1237.0	26.225899
Gintama.: Porori-hen	Manga	TV	13.0	8.59	23130.0	1464.0	450.0	26.123748
Gintama.: Shirogane no Tamashii-hen	Manga	TV	12.0	8.85	14981.0	1724.0	393.0	26.091383
Gintama: Yorinuki Gintama-san on Theater 2D	Manga	Movie	2.0	8.52	7014.0	2773.0	32.0	25.847349
Gintama°: Aizome Kaori-hen	Manga	OVA	2.0	8.50	16455.0	1979.0	87.0	25.871322
Hajime no Ippo: New Challenger	Manga	TV	26.0	8.73	77718.0	663.0	1074.0	28.093705
Hajime no Ippo: Rising	Manga	TV	25.0	8.66	59310.0	845.0	650.0	27.308071
Kamisama Hajimemashita: Kako-hen	Manga	OVA	4.0	8.54	27168.0	1305.0	517.0	25.586449
Kara no Kyoukai 7: Satsujin Kousatsu (Kou)	Light nove	Movie	1.0	8.52	66674.0	707.0	1446.0	28.242573
Kingdom 2nd Season	Manga	TV	39.0	8.51	21659.0	1621.0	503.0	27.432197
Kizumonogatari II: Nekketsu-hen	Light nove	Movie	1.0	8.69	71298.0	612.0	771.0	25.439327
Kizumonogatari III: Reiketsu-hen	Light nove	Movie	1.0	8.87	62314.0	649.0	1541.0	25.383845
Kuroshitsuji Movie: Book of the Atlantic	Manga	Movie	1.0	8.54	18459.0	1327.0	564.0	25.109284
Major S5	Manga	TV	25.0	8.54	23101.0	1835.0	435.0	28.120656
Mushishi Special: Hihamukage	Manga	Special	1.0	8.64	36043.0	1201.0	100.0	27.605896
Mushishi Zoku Shou 2nd Season	Manga	TV	10.0	8.83	48786.0	727.0	701.0	27.413434
Mushishi Zoku Shou: Odoro no Michi	Manga	Special	1.0	8.52	25401.0	1548.0	60.0	27.496842
Mushishi Zoku Shou: Suzu no Shizuku	Manga	Movie	1.0	8.71	22283.0	1432.0	124.0	26.988823

b. The second reason is that some of them belong to old anime and old genre. Military is not popular nowadays

ired_from_year	aired_from_month	premiered_season	first_genre	secc
1988	Jan	Spring	Military	

aired_from_year	aired_from_month	premiered_season	first_genre	se
1988	Jan	Spring	Military	

c. The third reason is second season

6 animes out of 37 high score but low popularity are the second season. For those people who love to watch second season of an anime, most users

are fans for the first season of that anime. It is harder to attract new audience for the second season than a brand new anime.

Conclusion

1. Anime of the hottest genre like comedy, adventure and action, do not have significant seasonal / monthly variance, maintain the popularity as always.
2. For Cars genre animes, spring may be in low response; For Police genre animes, potential market in Autumn; For Samurai genre animes, potential market in Winter
3. Animes adapted from games have potential.
4. Anime score and popularity are highly positively correlated. For exceptions, high score but low popularity, they are likely to be anime movie, old military anime or second season anime

Limitation

1. In this project, our datasets did not cover anime released starting from late 2018 till now, thus the analytics conclusions may fail to catch the trend of the recent two years.
2. In this project, we ignore some correlated factors that may affect the anime scores or users' behavior simultaneously. We assume that each feature affects the anime score or users' behavior separately.

Work Distribution

1. Peilin Zhong: Data Preprocessing, EDA, Question1
2. Zeyi Luo: Question2
3. Sijin Zhou: Question3

References

[1] https://en.wikipedia.org/wiki/Otaku

[2] https://myanimelist.net/

[3] https://www.kaggle.com/azathoth42/myanimelist

[4] https://myanimelist.net/topanime.php

[5] https://www.dictionary.com/browse/ecchi

[6] https://en.wiktionary.org/wiki/shoujo_ai

[7] https://en.wikipedia.org/wiki/Yaoi

[8] https://en.wikipedia.org/wiki/Yuri_(genre)

[9] https://www.dictionary.com/browse/harem?s=t

[10] https://en.wikipedia.org/wiki/Hentai#Classification

[11] https://en.wikipedia.org/wiki/Toei_Animation